Aalto University
School of Chemical
Technology

School of Chemical Technology
Degree Programme of Chemical Technology

Martina Ikonen

STUDYING SEQUENCE-EXPRESSION CORRELATION FOR RECOMBINANT PROTEINS

Master's thesis for the degree of Master of Science in Technology submitted for inspection, Espoo, 3 October, 2016.

Supervisor Professor Markus Linder

Instructor M. Sc. Bart Rooijakkers

Aalto University
School of Chemical
Technology

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi

Abstract of master's thesis

Author Martina Ikonen

Title of thesis Studying sequence-expression correlation for recombinant proteins

Department Department of Biotechnology and Chemical Technology

Professorship Biomolecular materials **Code of professorship** KE-30

Thesis supervisor Prof. Markus Linder

Thesis advisor(s) / Thesis examiner(s) M. Sc. Bart Rooijakkers

Date 3.10.2016 **Number of pages** 60 + 20 **Language** English

Abstract

In recombinant protein production the expression can be optimized on several levels, including the choice of cultivation parameters, the expression host and the expression system. However, the optimization of the open reading frame of the expressed gene is too often ignored, although it may have a drastic effect on the expression.

The optimization of the protein coding sequence does not have universal guidelines as little is understood about the contribution of different factors, and how to prioritize them. The key factors that affect the protein expression on gene level are GC-content, mRNA secondary structures, codon bias and rare codons, and the availability of charged tRNAs. However, these parameters are interdependent and changes in one affect the other, and they may have both local and global effects to the sequence.

Cellulose binding modules (CBM) are interesting in the biomolecular material science as they could be used in cross-linking and functionalizing cellulose. In this work the expression and function of CBM1 proteins were optimized and studied by shuffling two putatively homologous CBM1 sequences block-wise and so detecting the contribution of different parts of the sequence to the expression levels and functionalities. Additionally, the codon optimization was studied by comparing the wild type sequence to the optimized sequence of Cel7A CBM.

The CBMs were expressed as fusion proteins with alkaline phosphatase (AP) in several *Escherichia coli* expression strains. The expression levels were measured with AP enzymatic assay and sodium dodecyl sulfate gel electrophoresis (SDS-PAGE). The binding affinities of different CBMs were compared in binding experiments with nanosized cellulose and chitin.

The production yield of Cel7Aopt was 101 mg/l cultivation at the best, whereas the original Ehux1b2 was only expressed 38-56 mg/l. The replacement of C block by the Cel7A sequence made the Ehux1b2C to be expressed the same level as Cel7Aopt. CyDisCo strain increased the expression but also caused much variation between clones. However, the binding tests showed that the original Ehux1b2 did not have binding affinity towards cellulose, but the replacement of the D-block with the sequence from Cel7A recovered the functionality. No other shuffled protein showed any affinity towards cellulose, and well expressed Ehux1b2C bound even less to either substrate than the original Ehux1b2. Additionally, the sequence analysis of natural CBM1 proteins revealed that the D-block contained many amino acids which are conserved especially within cellulose binding modules, and which are presumably essential for the binding ability.

Keywords protein expression, *Escherichia coli*, protein shuffling, cellulose binding module, codon optimization, nanocellulose, chitin

Aalto-yliopisto
Kemian tekniikan korkeakoulu

Aalto-yliopisto, PL 11000, 00076 AALTO
www.aalto.fi

Diplomityön tiivistelmä

Tekijä Martina Ikonen

Työn nimi Sekvenssi-ekspressiokorrelaation tutkiminen rekombinanttiproteiineilla

Laitos Biotekniikan ja kemian tekniikan laitos

Professuuri Biomolekyylimateriaalit **Professuurikoodi** KE-30

Työn valvoja Prof. Markus Linder

Työn ohjaaja(t)/Työn tarkastaja(t) M. Sc. Bart Rooijakkers

Päivämäärä 3.10.2016 **Sivumäärä** 60 + 20 **Kieli** englanti

Tiivistelmä

Rekombinanttiproteiinien ekspressiota voidaan optimoida useilla eri tasoilla, kuten kasvatusolosuhteilla, sekä tuottoisännän ja ekspressiomekanismin valinnalla. Itse proteiinia koodaavan sekvenssin optimointi jää kuitenkin usein liian vähälle huomiolle, vaikka sillä voi olla suuri vaikutus ekspressioon.

Proteiinia koodaavan sekvenssin optimoinnille ei ole selkeitä sääntöjä, sillä geenitason eri tekijöiden vaikutuksista ekspressioon tiedetään hyvin vähän. Avainasemassa proteiiniekspression säätelemisessä ovat kuitenkin GC-pitoisuus, lähetti-RNA:n sekundäärirakenteet, kodonikäyttö ja harvinaiset kodonit, sekä aminohapollisten siirtäjä-RNA:iden saatavuus. Nämä tekijät ovat kuitenkin sidoksissa toisiinsa, joten yhden tekijän muuttaminen saa aikaan muutoksia myös toisessa tekijässä, ja vaikutukset voivat esiintyä paikallisesti tai koko sekvenssin tasolla.

Selluloosaan sitoutuvat proteiinit (cellulose binding module, CBM) ovat kiinnostavia biomolekyylimateriaalitutkimuksessa, sillä niiden kykyä tarttua selluloosaketjuun voitaisiin käyttää selluloosan funktionalisoinnissa ja luomaan verkkorakennetta. Tässä työssä CBM1-sekvenssejä tutkittiin sekoittamalla kahden CBM1-proteiinin sekvenssejä lohkoittain ja tutkimalla näiden neljän hybridiproteiinin avulla eri osioiden vaikutuksia proteiiniekspressioon ja proteiinien toiminnallisuuteen. Lisäksi kodonioptimointia tutkittiin ottamalla koeasetelmaan mukaan Cel7A-proteiinin villityyppisekvenssi optimoidun sekvenssin rinnalle.

Selluloosaan sitoutuvia proteiineja tuotettiin fuusioproteiinina alkaalifosfataasin (alkaline phosphatase, AP) kanssa useassa eri *Escherichia coli* -tuottokannassa. Ekspressiotasoja mitattiin alkaalifosfataasin entsyymiassaylla sekä natriumdodekyylisulfaattipolyakrylamidigeelielektroforeesilla (sodium dodecyl sulfate polyacrylamide gel electrophoresis, SDS-PAGE). Eri proteiinien toiminnallisuutta tutkittiin sitoutumiskokeilla nanokokoiseen selluloosaan ja kitiiniin.

Cel7Aopt-proteiinin saanto oli parhaimmillaan 101 mg/l kasvatusta, kun taas Ehux1b2 oli tuottotasoiltaan selvästi alhaisempi, vain 38-56 mg/l. Ehux1b2C, jossa C-lohko korvattiin Cel7A:n sekvenssillä, tuottui kuitenkin yhtä hyvin kuin verrokkina toiminut Cel7A. CyDisCo-kannan käyttö lisäsi proteiinin tuottoa, mutta aiheutti laajaa vaihtelua eri pesäkkeiden välillä. Alkuperäinen Ehux1b2 ei sitoutunut selluloosaan lähes lainkaan, mutta D-lohkon mutaatio Cel7A:n sekvenssin mukaiseksi teki proteiinin yhtä toimivaksi kuin verrokki. Muut substituutiot eivät lisänneet proteiinin sitoutumista, etenkään hyvin tuottunut Ehux1b2C. Lisäksi luonnossa esiintyvien selluloosaan tarttuvien proteiinien analyysi osoitti, että D-lohkossa on monia konservoituneita aminohappoja, jotka ovat oletettavasti tärkeitä juuri selluloosaan sitoutumisessa.

Avainsanat proteiiniekspressio, *Escherichia coli*, rekombinanttiproteiini, selluloosaan sitoutuvat proteiinit, kodonioptimointi, nanoselluloosa, kitiini

Forewords

This thesis was made as a part of the Academy of Finland's Centre of Excellence of Biosynthetic Hybrid Materials research, HYBER. The work was done in Aalto University School of Chemical Technology, supervised by Prof. Markus Linder.

I would like to thank my advisor M. Sc. Bart Rooijakkers who initially taught me the basics of hands-on protein work and later on was always ready to give valuable advice. Furthermore, I want to thank my supervisor, Prof. Markus Linder, as he had always new ideas how to improve the results, and what to do when the research turned up to be a dead end. Additionally, I want to thank the encouraging atmosphere in the biotechnology lab, and the general mindset to help everybody. Especially, I am grateful for having been a part of the brilliant Biomolecular Materials research group which already feels like a second family.

I also want to thank my dear chemist friends Minttu, Vilja, Wuokko, Tuomas, Tero and Katja for peer support throughout the studies and especially during the time we all were making our theses simultaneously. I also want to thank Laura for constantly exchanging thoughts and frustrations in the office and at lunchbreaks.

Finally, I am thankful for the support of my family throughout my whole studies. Vesa, Mom and Jussi, Dad and my best little siblings Patrick and Catharina, thanks for your encouragement!

Espoo, 21.9.2016

Martina Ikonen

INDEX

Abbreviations

AP	alkaline phosphatase
BC	bacterial cellulose
bp	base pair
CAI	codon adaptation index
CBM	carbohydrate binding module or cellulose binding module
ChNC	chitin nanocrystals
DNA	deoxyribonucleic acid
IPTG	isopropyl β-D-1-thiogalactopyranoside
kDa	kilodaltons (kg/mol)
mAU/min	milli absorbance units per minute
mRNA	messenger RNA
NFC	nanofibrillated cellulose
ORF	open reading frame
RBS	ribosomal binding site
RNA	ribonucleic acid
SDS-PAGE	sodium dodecyl sulfate polyacrylamide gel electrophoresis
tRNA	transfer RNA
V_0	the maximal rate of yellow color formation in the AP-assay

LITERATURE PART

1 Introduction

1.1 Protein expression optimization

Recombinant proteins are produced for a myriad of purposes, such as medical treatment, industrial use and biofuel production. To maximize the protein yield after downstream processing, such as extraction and purification, the protein expression in the cells needs to be maximized. The expression optimization can be divided in several levels (Figure 1, Gustafsson *et al.*, 2012). The growth conditions, including nutritional and temporal variables and choices of the culture volume, and downstream processing are generally carefully investigated and also adjusted on the go. The choice of host organism and strain depends on the application of the protein of interest, because different organisms modify the protein in different ways. The expression system, such as inducibility and copy number of the expression vector needs to be designed and chosen based on the nature of the protein of interest and targeted expression levels.

Usually relatively good care is taken of the upper level optimization. However, surprisingly little effort is put in the design of the open reading frame (ORF) itself, which has proven to affect the expression several hundredfold (H. Hu *et al.*, 2009; S. Hu *et al.*, 2013; Allert *et al.*, 2010; Gustafsson *et al.*, 2012). The ORF optimization has traditionally been mainly mimicking the codon usage of the host organism and avoiding rare codons (Welch *et al.*, 2011). Nearly every biotechnology company has their own algorithm for the optimization, but there is no public consensus on which properties are characteristic for a good sequence. However, the sequences can be additionally adjusted by selecting the codon usage in the sense of mRNA secondary structures, the strength of the nucleotide binding depending on GC content and adjusting the sequence to the codon usage of the host organism.

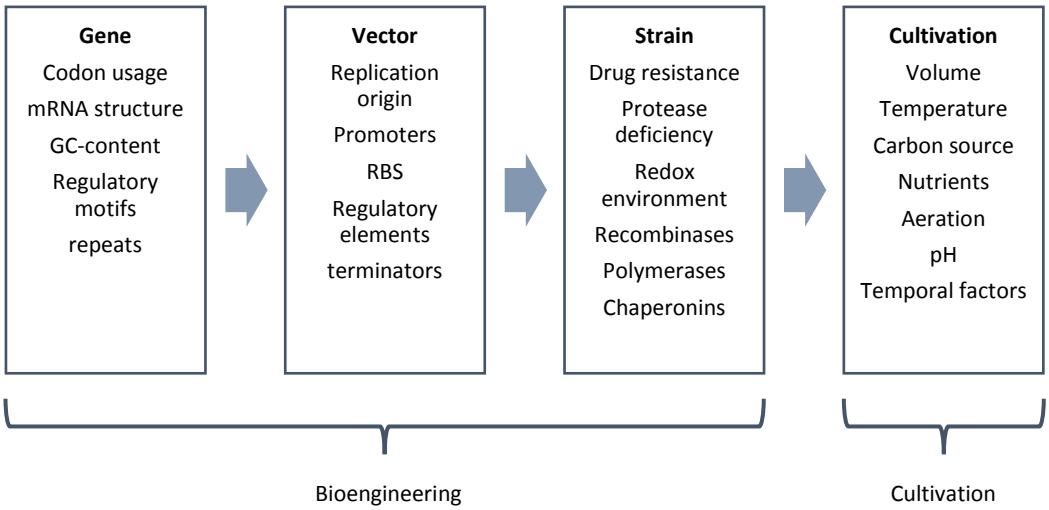

Figure 1. Different levels of protein production optimization and variables of each step. Bioengineering factors include modifications inside the cell, whereas cultivation factors are surrounding conditions that affect the growth of the cells. Image adapted from Gustafsson *et al.* (2012).

1.2 Cellulose and cellulose binding modules

Cellulose is predicted to be a material of the future. It is renewable and with its biodegradability it meets the requirements for a novel material to replace the oil-based polymers that are causing problems not being decomposable. Finland has also great raw material sources for cellulose, and the aim is to refine the cellulose material into more valuable products. However, to be able to use cellulose in wide variety of applications, it needs to be modified chemically or physically. The current ways to functionalize the nanocellulose surface are mainly chemical, so the hydroxyl groups are modified by covalent bonding. These reactions usually involve quite strong and toxic chemicals, such as strong acids and bases, solvents or reducing or oxidizing compounds. (Habibi, 2014; Arola, 2015; Abitbol *et al.*, 2016)

The nature gives inspiration to new kinds of biomaterials. For example a squid beak is one of the hardest non-mineral materials known, consisting of a composite

material made of chitin that is cross-linked with chitin binding proteins (Tan *et al.*, 2015). The high similarity of chitin and cellulose propose the possibility to do the same with cellulose. Proteins are elastic and functional by nature. Specific kind of proteins called cellulose binding modules (CBM) have natural ability to attach to cellulose by π-electron interactions. When combining the different features of proteins and cellulose, we could get cross-linked cellulose fibers with new functions. (Arola, 2015)

Carbohydrate binding modules (CBM) have been classified into several families based on their function and homology, and the CBM protein family 1 mainly consist of cellulose binding modules. On the other hand, the carbohydrate binding modules may have affinity towards several polysaccharides because of their structural similarity. Consequently, depending on the context, the acronym "CBM" refers to both words, cellulose binding modules more specifically, or carbohydrate binding modules in general. In the nature, cellulose binding modules are usually a moiety of cellulose degrading enzymes, where their function is to attach the enzyme to cellulose surface and thus bring together the substrate and the catalytic domain of the enzyme. Cel7A, the reference CBM used in this study, originates from the fungus *Trichoderma reesei*, and belongs to the CBM1 family (Palonen *et al.*, 1999). The binding properties of Cel7A have been attributed to 5 essential amino acids, especially three tyrosines which form the flat bottom surface of the CBM and bind to cellulose chains (Figure 2, Linder *et al.*, 1995).

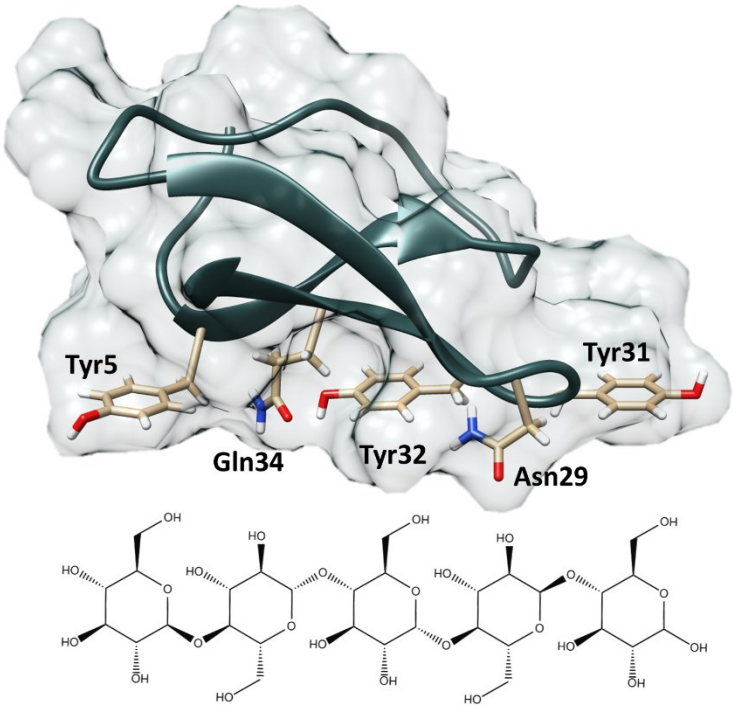

Figure 2. The structure of a cellulose binding module Cel7A (PDB ID: 1CBH, Kraulis *et al.*, 1989). The alignment with cellulose chain reveals the possible π-stacking of the aromatic residues to the sugar rings.

1.3 Protein homologues as a basis for expression optimization

Proteins whose function is predicted to be the same or closely related based on relatively high sequence similarity, but which may derive from different organisms, are called homologous proteins. Cellulose binding modules of protein family 1 are naturally found in many wood-degrading organisms such as filamentous fungi, like *Trichoderma reesei* (Martinez *et al.*, 2008). CBM1-like proteins have also been found in a coccolithophore *Emiliania huxleyi* (Figure 3), which is a unicellular marine phytoplankton species. *E. huxleyi* forms a calcite coccolith layer on top of its cell membrane, but otherwise it lacks cell wall. (Read *et al.*, 2013) In *E. huxleyi* genome, six putative homologues for *T. reesei* cellobiohydrolase 1 CBM are found in 5 different type of proteins. The existence of CBM1 genes in a marine organism is interesting and it suggests a structural function for the cellulose. In the coccolith formation compartments inside the cell membrane some polysaccharides have

been found, although their specific role in the calcification is not fully understood (Brownlee *et al.*, 2015).

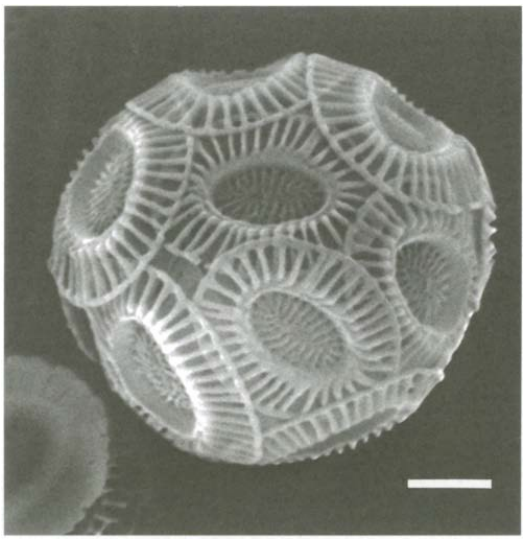

Figure 3. Coccolithophore *Emiliania huxleyi* in scanning electron micrograph. Scale bar 1µm (Paasche, 2001).

As previously observed (Ikonen, unpublished data), the expression levels of different protein homologues may vary significantly. Shuffling of different homologous sequences is a way to study both codon optimization for maximal expression as well as the structure and function of the protein. When homologous sequences are shuffled, the structure and fold of the protein can be mostly maintained but local variations can be tested. Consequently, amino acid level forms one more level of optimization, although the key to the expression variation may lie in nucleic acid level caused by different protein sequences.

1.4 Overview of the experimental part

The objective of the experimental part was i) to optimize the CBM1 protein production in *E. coli* and ii) to investigate the differences of the expression and function of the homologous CBM1 sequences. In the experimental work, 7 different sequences for CBM of family 1 were compared. The reference CBM1 sequence (Cel7A, PDB ID: 1CBH) originates from the filamentous fungus

Trichoderma reesei cellobiohydrolase 1. In the pan genome of *Emiliania huxleyi* coccolithophore species six putatively homologous CBM1 protein sequences have been found (Read *et al.*, 2013), and one of them (UniProtKB: R1C3S1, position 325-361, here named as Ehux1b) was compared to the reference sequence in this thesis. To chart the contributions of different parts of the sequence to the protein properties, four artificial variants were made by shuffling the sequences block-wise. In addition, the effect of codon optimization was investigated by comparing the codon optimized and wild type DNA sequences of Cel7A.

Although in this thesis the target was primarily to study how the sequence shuffling affects the expression levels, a bigger expression level is of no use, if the protein does not work for its purpose. Consequently, the binding properties of each protein were compared with three different substrates, nanofibrillated cellulose (NFC), bacterial cellulose (BC) and chitin nanocrystals (ChNC). The binding of different variants were compared to learn about the function and mechanism of binding.

The following chapters describe different intracellular ways to optimize protein expression with the focus on the methods used in the experimental part of this thesis. The experimental part of the thesis gives a cross-section of the issues of protein production, concentrating especially on sequence-specific factors.

2 Strain engineering

2.1 Overview to strain engineering

One way to optimize the protein yield is to maximize the production of the protein per cell. It obviously causes more stress for the cell, so it needs a strain, which rather produces protein than biomass. Enhancing the protein expression level stresses the cell and may result in poor growth and thus in the end reduced

expression (Hu *et al.*, 2009). Another point of view is not to try to exhaust the cells by extensive production but to try to achieve higher cell densities in the specific culture volume. In order to make the strain suitable for any other purpose, like using different nutrients or deleting unwanted pathways that would interfere with the ones designed, the solution is to engineer the strain. The strain engineering defines the basics how the cell behaves and what kind of strategy will be used in the optimization.

The abundancy of sequence information of whole genomes of different organisms and strains enables detailed tuning of microbial strains. Instead of the previous approach of trial and error in random mutations and screening of the best clone, the mutations can be targeted to a specific gene, whether the aim is to boost it or to delete it. To be able to make sensible changes in the genome, it requires information about the gene functions, which can in turn be estimated by searching for functions of its homologies in other organisms. (Gustafsson *et al.*, 2012) The strain engineering also requires deep understanding of the cell metabolism, in order to make a strain that does not suffer from the modification too much. As traditional gene editing technologies have been time-consuming and troublesome, usually in protein production the strain is chosen from the selection of well-known and widely used industrial or laboratory strains, and the expression is tuned some other way. However, as the new genome editing technologies such as CRISPR-Cas9 and CRISPR-Cpf1 are continuously studied (Jinek *et al.*, 2012; Cong *et al.*, 2013; Zetsche *et al.*, 2015; Lander, 2016), the knowledge of genome editing is becoming abundant and the effort needed to edit the genome is getting much smaller. The reliable and more precise technologies enable quicker and less challenging genome modifications. Being cost- and time-effective, genome editing could be more widely utilized technology in the optimization of protein production. Creating synthetic pathways in organisms also requires efficient technologies to modify the genome itself instead of just plasmid-based systems, which are less stable.

2.2 Examples of *E. coli* production strains

E. coli BL21 is one of the most widely used laboratory expression strains. DE3-derivative of BL21 enables the use of T7 RNA polymerase based expression systems. Like every strain, BL21 also has its disadvantages, which has resulted in several attempts to modify the strain. T7 Express is one of the derivatives, which aims to be more stable in stressful growth conditions, maintain better foreign DNA, and be more suitable for the expression of DNA-active enzymes (E.A. Raleigh, personal communication, 27.6.2016). BL21 and T7 Express both have all the modifications in their genomes, but another way to modify a strain is to introduce an external plasmid, which contains additional genes that alter the metabolism. CyDisCo technology (Matos *et al.*, 2014) is targeted to outcompete the previous disulfide bond forming strains that were made by interfering the essential genes in the genomes. As will be noticed in the following examples, the strain tuning is dependent on very slight modifications and the expression levels of single enzymes. Moreover, the improvements are always made to some specific situations and there does not exist a universal perfectly working strain, but different strains need to be chosen for each purpose. As discussed before, the fine-tuning of commercial strains may increase if the CRISPR-technologies prove to be practical.

2.2.1 BL21 (DE3)

BL21 is an *E. coli* B strain, and a widely used laboratory expression strain (Studier and Moffatt, 1986). B strains are made deficient in *Lon* and *OmpT* outer cell membrane proteases in order to reach higher protein yields without degrading the protein at the purification (Grodberg and Dunn, 1988). On the other hand, as little protein is degraded, misfolded proteins may accumulate, so in case of problems in the translation or protein folding, resources of the cell are wasted.

BL21 (DE3) has an additional λDE3 prophage integrated into the genome. The prophage encodes for T7 RNA polymerase (*T7 gene 1*), which is controlled by *lacUV5* promoter (Studier, 2005). The gene of interest will be placed in a plasmid under a T7 promoter region, and the genomic T7 RNA polymerase can start transcribing the gene upon induction by isopropyl β-D-1-thiogalactopyranoside (IPTG) or lactose, which activates the *lacUV5* promoter. The simplicity of the T7 expression system has made it almost a standard for recombinant protein production. Nevertheless, DNA damage in the cells activates a pathway called SOS cascade, which is meant to correct the damage by homologous recombination and allow mutagenesis and evolution (Alberts *et al.*, 2008). However, in case of DNA damage, the SOS response activates the viable prophage λDE3, which in turn results in the cell lysis. Another disadvantage of the λDE3 strains is that the highly active T7 RNA polymerase will transcribe the gene of interest even uninduced (Studier, 2005). In the case of toxic protein expression, this may be fatal to the cell growth, and the transformation of toxic protein containing plasmids may not be possible. In case of possible toxic proteins the use of LacI repressor is vital and addition of T7 lysozyme needs to be considered, as it cuts out the basal expression (Studier, 1991). For these cases there are naturally other BL21 strain derivatives commercially available. Additionally BL21(DE3) has an increased resistance towards a virulent bacteriophage T1 (New England Biolabs, 2016a).

2.2.2 T7 Express

T7 Express (C2566 or ER2566 in New England Biolabs publications) is an enhanced version of the traditional BL21(DE3) expression strain. To overcome the problems of BL21(DE3), several modifications have been made in the genome (New England Biolabs, 2016b). First of all, in T7 Express the genomic copy of the gene for T7 RNA polymerase (T7 gene 1) is located in the lac operon instead of being in the prophage. The removal of the λDE3 and integration of only the T7 gene1 disables the lysis function and the strain is more stable in stressed conditions (E.A. Raleigh, personal communication, 27.6.2016). As T7 Express strain also lack the Lon and

OmpT proteases, its capability of degrading any proteins is compromised. A cell division inhibitor protein SulA, which is expressed as a result of the SOS cascade, would thus accumulate after SOS response, preventing the cell division. T7 Express has a mutation in the SulA promoter, which diminishes the expression of SulA and makes the strain less sensitive to the consequences of the SOS response (New England Biolabs, 2016b). As the T7 Express strain is created mainly for better production of DNA active enzymes, such as restriction endonucleases and DNA binding proteins, several interfering elements have been removed from its genome. To stabilize foreign DNA in the cells the genes of methylation and restriction enzymes have been mutated or deleted. (E.A. Raleigh, personal communication, 27.6.2016) This way the other DNA active enzymes will no longer exist in the cytoplasm, which also makes the purification of the commercial protein of interest easier.

2.2.3 Enhancing disulfide bond formation in *E. coli*

An abundant bottleneck for protein production is the misfolding of the protein and the accumulation of the protein to inclusion bodies in the insoluble fraction. Especially in the case of heterologous production the protein of interest is likely to misfold, or to fold incompletely, as the folding machinery is not optimal for the protein. The disulfide bonds cannot be folded efficiently in the cell cytoplasm because of the reducing environment. Unlike eukaryotes, prokaryotes like *E. coli* also lack the endoplasmic reticulum, which would have a reducing environment to assist the disulfide formation and correct folding. (de Marco, 2009) Thus heterologously produced protein, which naturally folds via disulfide bond formation in the post-translational modification pathway, is presumably misfolded and not functional.

Several different approaches have been taken to overcome this issue. The eukaryotic production, such as with yeast, has its own limitations, like higher production costs and hyperglycosylation of the proteins. The genetic modification

of *E. coli* to enhance disulfide bond formation is the other commonly used way, which has resulted in new commercial production strains, like Origami strains (Novagen). Origami strains have mutations in the key enzymes of the glutaredoxin system (*gor/TrxB*), which normally maintain the reducing environment in the cell cytoplasm. However, interfering the redox potential of the cell affects negatively the growth of the cells, and the doubling time in regular growth medium is as much as 300 min (Prinz *et al.*, 1997; Åslund and Beckwith, 1999). As presumably every gene has its own contribution to a living organism, mutations in the genome unbalance the system. Therefore, adding genes rather than deleting any might be safer choice in some occasions. If the normal cell metabolism is not interfered and the growth rate is maintained relatively high, the overall production is better even if there are many resource-consuming genes to be expressed.

CyDisCo technology (Cytoplasmic disulfide bond formation in *E. coli*) is a plasmid-based technology, so it can be used in any strain, like previously mentioned BL21(DE3) and T7 Express. CyDisCo enhances the protein folding by catalyzing the disulfide bond formation. The technology comprises several plasmids with slight differences, but two enzymes form the core functions: sulfhydryl oxidase (Erv1p) catalyzes the disulfide formation whereas protein disulfide isomerase (PDI) catalyzes the disulfide bond isomerization (Matos *et al.*, 2014). The CyDisCo version used in the experimental part of this thesis is pMJS205, which consists of the genes for Erv1p and PDI, p15A origin of replication and chloramphenicol resistance marker (Figure 4, full sequence available in the Appendix 1). The CyDisCo components, Erv1p and PDI are controlled by *tac* promoter, an efficient artificial promoter which can be repressed by the LacI repressor and activated by IPTG (de Boer *et al.*, 1983). In this study the LacI is encoded in the sequence of the other expression plasmid, which contains the protein of interest. Consequently, the CyDisCo plasmid is not efficiently repressed in cases where it is the only plasmid in the cell.

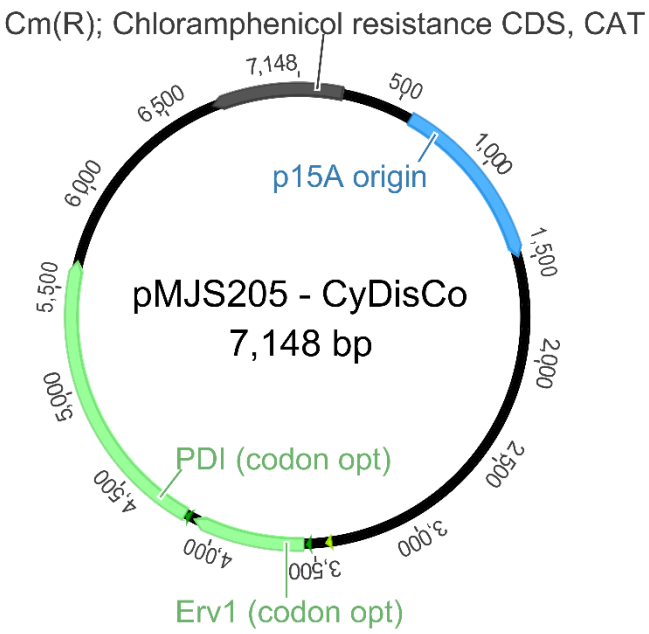

Figure 4. CyDisCo plasmid pMJS205 contains genes for Erv1p and PDI, and p15A origin of replication and chloramphenicol resistance marker (Image by *Geneious*, Kearse *et al.*, 2012).

3 Expression systems

3.1 Vector engineering

The vectors that we use nowadays utilize known parts of natural organisms and their genetic pathways. However, the expression vector selection is not yet well categorized, and sequence level tuning of vectors is not in the priority. The vectors that are currently in use are constructed from natural regulatory elements, but the elements itself are not specifically optimized for the production. (Gustafsson et al., 2012) There have been many attempts to standardize and classify different elements of expression vectors, such as replication origins, selection markers and regulatory sites like promoters and terminators (Shetty et al., 2008; Silva-Rocha et al., 2013). There are also some studies about computational analyses which aim

to better classify the regulatory elements (Jonsson *et al.*, 1993; Omotajo *et al.*, 2015).

Databases containing detailed information about the properties of each component are useful when creating novel expression systems. As the *de novo* DNA synthesis becomes affordable to more and more researchers and less cloning is performed by hand, the information from such databases can be used when ordering custom made vectors. The ready-made vectors that several companies offer may remain a good starting point but supposedly will sooner or later become impractical.

3.2 T7 expression system

T7 expression system is based on a genomic T7 RNA polymerase, which upon induction can transcribe the gene cloned downstream of a T7 promoter region. Usually the T7 RNA polymerase gene *T7 gene 1* is regulated by lactose or its analogue IPTG, or some other sugar such as arabinose. The T7 system also enables multiple genes being activated by the induction of T7 polymerase. The strains that have a mutation in the lactose permease gene *lacY1* additionally enable the tuning of the expression level. The mutation allows uniform intake of IPTG in the cells, and thus the expression levels can be adjusted precisely. Lower expression rates may allow better folding for some proteins and that way increase the amount of functional protein (Novagen, 2006)

The wide usage of T7 RNA polymerase system in recombinant protein expression stems from the nature of the polymerase itself. It is highly selective in producing the protein coded under T7 promoter regions, and the speed of T7 RNA polymerase can be 5 times higher than the *E. coli* natural RNA polymerases. In nature, this is an advantageous virulence property as the T7 bacteriophage has a way to direct the whole cell's energy to transcribe only its own genes. This results in a situation where the protein of interest may constitute even half of the protein

production of the cell. However, as the polymerases do not compete for the same templates, the effectiveness of T7 polymerase lies in its speed as it steals ribonucleotide construction material from other slower RNA polymerases. (Studier and Moffatt, 1986) The high selectivity of the promoter regions disables almost any expression in the absence of the T7 RNA polymerase, which is a good property in the case of a toxic protein overexpression (E.A. Raleigh, personal communication, 27.6.2016).

4 ORF optimization

4.1 The paradox of ORF optimization

Due to the advances in the *de novo* DNA synthesis and the reduction of prices, it is easier to synthesize a new DNA sequence than to extract it from the native species as was still being done just a few years ago. As in the DNA synthesis virtually any sequence can be synthesized, the genes are usually optimized for the expression host. The challenge lies in the optimization, as there are no universal guidelines or knowledge how it should be done in order to maximize the expression. Having no thumb rule for the optimization, the different sequences usually simply need to be tested.

The lack of reliable design rules to create a good protein coding sequence derives from the fact that the gene level biology cannot yet be predicted and simulated. To be able to simulate and predict the behavior, one should have plenty of data, based on which the simulation could be built. The difficulty in the gene level experiments is that testing everything in the laboratory consumes a lot of time, money and raw material resources, and all combinations are not even possible to be tested without quite heavy automatization. However, the research can be started for example by searching among existing data, like bacterial genomes, and

compare the properties and behavior of different parts of the transcripts (Allert *et al.*, 2010).

To make the optimization even more confusing, the chemistry of DNA also contains other information in addition to the basic amino acid sequence information: transposon resistance, mRNA processing and folding, DNA folding and packing, and RNAi regulation. These properties have not been intentionally coded in the DNA like nothing in the nature, but they have survived from the natural selection being advantageous to the host organism. (Gustafsson *et al.*, 2012) As not all reasons for the advantageous properties are known, it makes it hard to create such multilayered information in artificial genes in just a single type of code.

The optimization of ORF can be roughly divided into two categories: the optimization of transcription and the optimization of translation. There is also a possible third level – the effect of amino acid substitutions to the protein and its expression. Nevertheless, the amino acid level optimization is rather optimizing the function of the protein, and with the substitutions to the amino acid sequence, it also affects the possible choices for a successful DNA sequence. The third level is of the major interest in the experimental part of this thesis.

4.2 Studying the ORF optimization

The simulations require experimental data about how different sequences perform. Nowadays the shuffling can be performed in silico, and then synthesize the ready genes, although being creative helps to achieve much more variation in the sequence just in few steps. Even though gene synthesis is becoming cheaper, the power of natural recombination still outcompetes the synthesis resources in the variation. As automation is taking place in the lab, more and more screening can be performed in a shorter time. However, as seen in the study of Hu *et al.* (2009) and depending of the resources available, the limiting step for the research

can also be the sequencing, which would only be done for the few variants selected for further study.

Nowadays DNA synthesis is highly efficient and readily available, and it allows the creation of quite many sequences in competitive price. Allert *et al.*, (2010) simply exploited the power of *de novo* gene synthesis and created and compared 285 genes having synonymous codon usage. If there is no chance of synthesizing all the wanted variants, the choices are to modify DNA somehow. Error-prone PCR as well as digestion and reassembly of a gene are methods, which can cause single point mutations in the genes. However, the rate of getting new properties might be quite slow and requires many cycles (Stemmer, 1994a, 1994b; Moore and Arnold, 1996; Crameri *et al.*, 1998). As the mutations can be anywhere and additionally cause amino acid substitutions, the point mutations are better suited for the directed evolution of proteins rather than studying the codon usage of a single amino acid sequence. Making point mutations is also a quite old technology, and it was more widely used when the genes were obtained from the native species instead of synthesis, and the whole genes could not be rewritten. Crameri *et al.*, (1998) compared point mutation versus multi-gene shuffling in improving the protein, and the multi-gene shuffling resulted at the best in 540-fold enhancement in the activity, whereas point mutations gave only up to 8-fold increase in function.

One way to introduce more variation in the sequence is to create sequences from oligonucleotides, which are slightly different (Hu *et al.*, 2013). As the oligonucleotides recombine different ways, the ultimate number of different sequences is enormous. The mutations in the primers may be designed so that they will only cause synonymous codon changes, i.e. the amino acid sequence will remain the same, and the codon optimization can be tested. Kudla *et al.* (2009) used a combination of two methods in order to further increase the variation in the sequences. They created fragments of the ORF sequence by PCR from oligonucleotide mixes, which contain synonymous mutations at third positions.

After sequencing the fragments, they were digested from the ends in order to ligate them together as different combinations. (Kudla *et al.*, 2009) Two steps of different combinations causes exponentially growing amount of variants. Varying the third codon positions is the simplest way to introduce synonymous mutations, as many amino acids are encoded by codons that differ only by the third position.

Family shuffling is shuffling between the genes of proteins that belong to the same protein family, but originate from the same or different species. Just by digesting DNA of homologous genes with randomly digesting restriction enzyme DpnI or DNase and then reassembling with recombination in primerless PCR, countless amount of variants can be created (Stemmer, 1994a; Crameri *et al.*, 1998; Hu *et al.*, 2009). Family shuffling usually includes slightly different amino acid sequences, so the amino acid usage is usually also changed, and thus the changes in the function of the protein may be significant. However not all mutations in the protein sequence cause major differences in the functionality but the nucleotide difference may affect the expression rate.

Although massive variation can be achieved using various techniques based on random mutations or combinations, the shuffling can also be made by using rational design. Welch *et al.* (2009a) studied the gene expression of synthetic segmentwise shuffled genes. They compared two sequences that differed by the synonymous codon usage, and made three chimeras having different combinations of the fragments. By detecting the contribution of each fragment, the properties of advantageous or on the other hand deleterious fragments can be further studied.

4.3 Properties of a good ORF sequence

Several factors affect the gene expression on ORF level: messenger RNA levels affect the translation rate, as the more there is available RNA, the more strands ribosomes can start translating. The amount of mRNA available is in turn

dependent of the variables of transcription, such as CG-content of the sequence. The velocity of translation is fundamentally determined by the availability of charged transfer RNAs (tRNA) - the rare codons are codons which are not frequently used in the organism and consequently there are very few or no charged tRNAs available for such codons. The initiation of translation is a critical phase, and problems in at the initiation can easily result in deficient expression. (Gustafsson *et al.*, 2012) Moreover, the sequence codon choices can affect locally or globally. In local prevalence, the codon choices may cause rare codon clusters or mRNA secondary structures that may prevent the translation or cause premature termination. When distributed throughout the sequence, the codon bias is the slowing down the translation rate as not enough tRNAs of specific codons are available. (Welch *et al.*, 2009a) As the codon choices have both local and global effects, the distinguishing of their effects is difficult.

The codon bias of the recombinant gene has been under debate since its discovery as there is no overall consensus about its effects (Gustafsson *et al.*, 2012). One predominant assumption has been that the codon usage of the recombinant gene should be biased like the codon usage of the host organism, especially mimicking the bias in the highly expressed genes of the host. This dates back to the research of Ikemura (Ikemura, 1982), who compared the codon usage of the genes to the availability of the tRNA molecules. Later Sharp and Li (Sharp and Li, 1987) proposed the *Codon adaptation index* (CAI) as the measure of the adjustment of the gene to the codon usage of the host. Value 1 is given to the codon used most often, so a gene using only the most frequent codons has a CAI value of 1. Whether the gene should only use the most frequent codons, or if the codon usage ratio should match the natural ratio, has been under discussion for decades (Welch *et al.*, 2009b).

Kudla *et al.* (2009) made experiments with 154 synonymous sequences of green fluorescent protein, but they found no effect of codon bias on the expression. Neither the frequency of the highest expressed codons nor the CAI showed any

correlation to the results. This suggests that rather than statistically choosing the right ratio of different codons, more important are the specific points, which may cause mRNA secondary structures or something else that hinders the translation significantly. However, the CAI did affect the cell growth, as the availability of all tRNAs affects the general fitness of the cells and possibly reduces the amount of misfolded proteins that need degradation. (Kudla *et al.*, 2009) On the other hand, the codons that are most favored in the expression were found to be those that stay charged the best during the starvation in the cell (Welch *et al.*, 2009b). Although the CAI itself did not correlate strongly with the expression, the CAI might be a good thing to keep in mind and use as secondary optimization factor.

The so-called rare codons, which are used in the own genes of the host organism significantly less frequently than some others, polarize opinions as some would like to avoid them totally and others would bias their usage according to the natural bias. When positioned in a cluster they might cause serious problems like prevent the whole translation by premature termination or cause frameshift at the ribosome (Gustafsson *et al.*, 2012). To some extent the synthesis of specific tRNAs have been enhanced by introducing genes for the rare codons needed into the strain (Novagen, 2006). Nevertheless, their use for the research has been questioned as the rule for the optimization is not clear and some rare codons are expressed well (Gustafsson *et al.*, 2012). It has also been proposed, that having slowly translated rare codons at the domain boundaries or other important positions might help the folding of the protein by giving it a small break to fold properly before continuing to the next domain (Angov *et al.*, 2008; Tsai *et al.*, 2008). In prokaryotic protein synthesis the protein starts to fold immediately from the N-terminus while it is still being synthesized at the ribosome. As the protein rushes to folding, there is a risk of deficient folding or misfolding. Since the qualities and the production yields of proteins made from different synonymous genes have varied significantly (Allert *et al.*, 2010; Hu *et al.*, 2013), the codon usage must be the key to the explanation. However, no conclusion could be made about which codon mutations or features are critical and have the most effect.

The ratio of A-T and G-C nucleotides affect how tightly the nucleic acid is bound as a double strand. Each G-C base pair has three hydrogen bonds whereas A-T base pairs only have two, so more energy is needed to open a G-C bond. Therefore the GC content must be adjusted to the preferences of the host organism. Long stretches of AT nucleotides may result in too weak binding and premature detachment of the strands, and for example terminate the transcription too early (Gustafsson *et al.*, 2012). On the other hand, high GC content especially in the ends of the transcript may be deleterious, as it would make the possible hairpin structures more stable (Allert *et al.*, 2010). Messenger RNA can fold into secondary hairpin structures after transcription. If the hairpin is strong, the ribosome may not be able to bind the ribosomal binding site (RBS). Therefore generally strong mRNA secondary structures are avoided in the sequence. Although less GC nucleotides make the mRNA secondary structures less stable, AT-rich sequences have been reported to be poorly expressed in *E. coli* for unknown reason (Plotkin and Kudla, 2011).

Allert *et al*., (2010) analyzed the open reading frames of over 800 bacterial genomes, and found out a tendency of less mRNA secondary structure in both ends of the transcript, especially in the 5’ end. In both ends of the mRNA the content of GC nucleotides was statistically lower indicating a lower energy to form any such 3D-structures. The codon adaptation index (CAI) values suggest that there were also slightly more rare codons in the beginning of the transcript. However as many less used codons are AT-rich, their contribution to the lesser mRNA-structure and higher AT-content of the 5’ end of the sequence can be seen as lowered CAI. (Allert *et al.*, 2010) Kudla *et al.* (2009) analyzed the expression of synonymous genes and found out the same correlation between the expression levels and mRNA structure at the 5’ end. However, Hu *et al.* (2013) noticed that the 5’ sequence does not exclusively explain the yield but the following codon choices have equal possibility to affect the expression. In both studies the expression levels as well as the functionality of the protein were greatly affected by the synonymous codon choices (Kudla *et al.*, 2009; Hu *et al.*, 2013). As a

conclusion, strong mRNA secondary structure should be avoided especially in the 5’ end of the ORF, but also later if possible.

Despite the lack of knowledge of the recipe for a successful gene sequence, there are some sequence motifs that are generally regarded as deleterious and should be avoided. RNase cleavage sites and transcriptional terminators are good examples what kind of sequences might cause problems to the transcription as those sequences clearly have a role to destroy or terminate the mRNA. (Gustafsson *et al.*, 2012) When creating the sequence, the susceptibility of such motifs need to be taken into account, and kept in mind. When optimizing the sequence following other design rules, there is a risk to accidentally include such motifs.

EXPERIMENTAL PART

5 Materials and methods

5.1 Creating of the sequences *in silico*

To study the differences in expression of highly similar homologous proteins, a set of variants were made using *Geneious* version 8.1.8 software (www.geneious.com, Kearse *et al.*, 2012). First the codon optimized CBM sequences of one *Emiliania huxleyi* CBM1, Ehux1b (UniProtKB: R1C3S1, position 325-361, Read *et al.*, 2013), and Cel7A (Malho *et al.*, 2015) were aligned and Ehux1b was truncated to match the length of Cel7A. The codon usage in matching amino acids was taken from Cel7A optimized sequence, and the new sequence was called Ehux1b2. The sequences of Cel7A and Ehux1b2 were divided in four blocks so that each block had 4-6 differing amino acids between Cel7A and Ehux1b2 (Figure 5). Four chimeric proteins were made by shuffling the blocks. Each chimeric protein contained one block from Cel7A and three blocks from Ehux1b2 (complete DNA sequences are available in the Appendix 1). The sequences supplemented with appropriate cloning sites were purchased as GeneArt strings DNA fragments from Life Technologies.

Figure 5. The sequences of CBMs compared in this thesis. The sequences of Cel7A and Ehux1b2 were divided in four blocks, which then were shuffled yielding 4 chimeric proteins, each with one block of Cel7A CBM and the rest three blocks from Ehux1b2. The coloured background of residues indicates equivalence of the sequence compared to Cel7A. (Image by *Geneious*, Kearse *et al.*, 2012)

5.2 Designing the expression principle

To make the production easier and to be able to detect easily the expressed protein, the CBM sequences were expressed as fusion proteins with alkaline phosphatase (AP) (Figure 6), which is well expressed in *E. coli*. Additional pelB signal sequence directed the folding protein into the periplasmic space of the cell to help in the folding and disulfide bond formation, and the C-terminal polyhistidine tag enabled the purification with affinity chromatography. The linker between alkaline phosphatase and the CBM contained several sites for trypsin cleavage in order to extract only the CBM parts if needed. The fusion proteins were expressed in pET28-derived pBR1a vector (Figure 7).

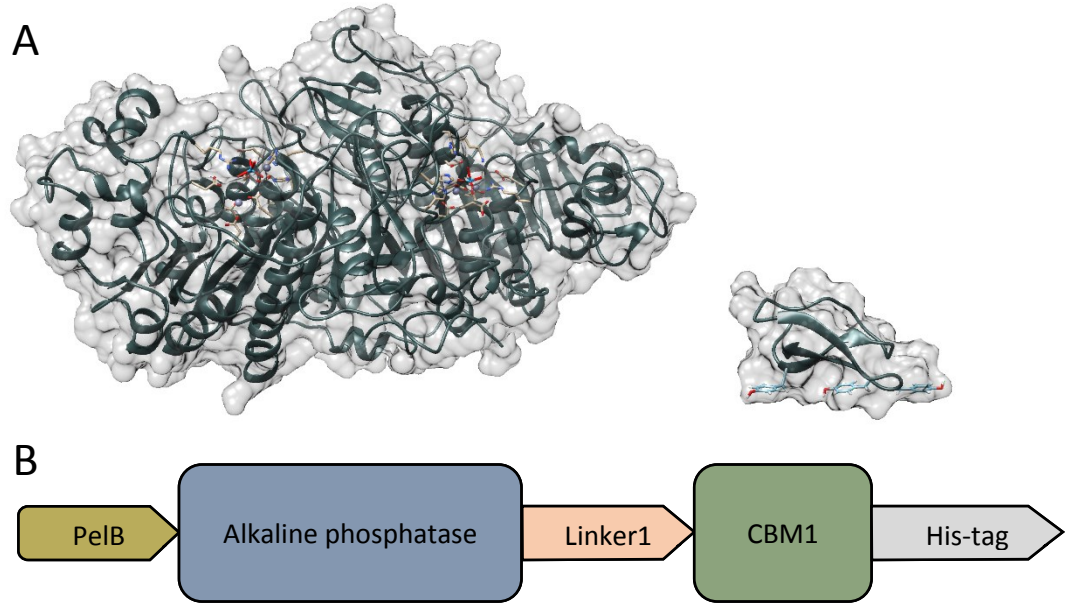

Figure 6. A) The illustration of alkaline phosphatase (PDB ID: 5C66, left) and Cel7A CBM1 (PDB ID: 1CBH, right) proteins (images made with UCSF Chimera). B) The representation of the AP-CBM-construct.

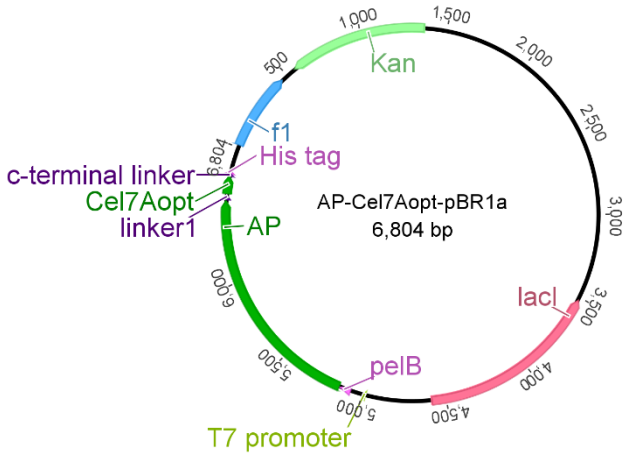

Figure 7. The plasmid map of the pET28-derived expression plasmid (Image by *Geneious*, Kearse *et al.*, 2012).

5.3 Golden Gate cloning

The Golden Gate cloning (Engler *et al.*, 2008, 2009) was performed using the following protocol: 50 ng of vector DNA, 50 ng alkaline phosphatase fragment DNA, 60 ng pelB-linker fragment DNA (amplified with primers MI02&MI03 following

KAPA HiFi Hotstart PCR kit instructions and purified with NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel)), 25 ng CBM DNA, 1 µl Cutsmart buffer (New England Biolabs), 10 mmol ATP (diluted from ThermoFisher R0441), 0,3 µl BsaI-HF (New England Biolabs), and 0,5 µl T4 DNA ligase (New England Biolabs) were mixed and sterile Milli-Q water was added to total volume of 10 µl. The control had vector DNA, Cutsmart buffer, ATP and BsaI-HF and the reaction volume was filled to 10 µl with sterile Milli-Q water. The reaction followed the thermocycle of 5 min in 37 °C and 5min in 16 °C (22 repeats), final digestion of 40min in 50 °C and inactivation for 10min in 80 °C. The reactions were transformed into chemically competent Top10 *E. coli* cloning strain. Plasmid extraction was performed with NucleoSpin Plasmid kit (Macherey-Nagel).

5.4 Transformation - chemically competent cells

Chemically competent cells were prepared by growing 100 ml 1:100 diluted LB medium culture to OD_{600} of 0,25-0,3 in 37 °C, 250 rpm. The culture was chilled and the medium was separated by centrifugation in 3200 x g, 4 °C, for 10 min. The cells were resuspended in 0,1 M $CaCl_2$ solution and incubated on ice for 30 min. After centrifugation in 3200 x g, 4 °C, for 10 min, the supernatant was removed and the cells were resuspended in 0,1 M $CaCl_2$ and 15 % glycerol solution and aliquoted for snap freezing with liquid N_2 and storing in -80 °C.

In the cloning *E. coli* Top10 strain was used and in expression BL21(DE3) and T7 Express strains were used. 50 µl $CaCl_2$-chemically competent cells were thawed on ice and mixed with 10 ng of plasmid DNA (100 ng of each plasmid in double transformations) or 5 µl of Golden Gate reaction mixture and let to stand on ice for 5 min. The cells were put in 42 °C water bath for 45 seconds and then again incubated on ice for 2 min, after which 250 µl SOC medium was added. The cells were let to recover in 37 °C incubator, for 30-60 minutes (depending on antibiotic resistance). Appropriate dilutions were made and the cells were plated on LB agar

plates, containing appropriate antibiotics (50 µg/ml kanamycin and/or 50 µg/ml chloramphenicol), and the plates were incubated in 37 °C overnight.

5.5 Transformation - electrocompetent cells

Electrocompetent cells were prepared by growing 2x150 ml 1:100 diluted LB medium culture to OD_{600} of 0,5-0,6 in 37 °C, 250 rpm. The culture was chilled and aliquoted into eight 50 ml Falcon tubes. All the following steps were carried out on ice. The tubes were centrifuged 3200 x g in 4 °C for 10 min, and the supernatant was replaced with 30 ml ice cold MQ-water. The centrifugation was repeated and the cells were resuspended in 15 ml MQ-water and two tubes were combined in order to have four tubes left. The previous step was repeated to get two tubes. In the final step, the two tubes were centrifuged and the cells were resuspended in about 1 ml of ice cold 50 % glycerol and aliquoted. The tubes were snap frozen with liquid N_2 and stored in -80 °C.

A tube of 40 µl electrocompetent cells was thawed and 1 µl DNA (10 ng) was mixed with the cells in cooled tubes and then incubated on ice for 5 min. The mixture was transferred to pre-cooled 2 mm cuvettes and 2,5 V heat shock was given (about 5,20 ms), followed by addition of 960 µl SOC medium. The cells were let to recover on ice for 2 min, and then in 37 °C for 1h. Appropriate dilutions were made and the cells were plated on LB agar plates, containing appropriate antibiotics (50 µg/ml kanamycin and/or 50 µg/ml chloramphenicol), and the plates were incubated in 37 °C overnight.

5.6 Colony PCR

Colony PCR was performed following the protocol of KAPA2G Robust HotStart ReadyMix PCR Kit (Kapa Biosystems). The AP-CBM constructs were amplified using pSVAR1 and pSVAF2 primers.

5.7 DNA electrophoresis

The colony PCR reactions were analyzed with DNA gel electrophoresis using 1% UltraPure Agarose (ThermoFisher 16500-100) in 1x TAE buffer. The 50 ml gels were stained at casting with 5 µl 10000x SYBR™ Safe Stain (ThermoFisher). The gels were run with 80-130 V as long as needed for the separation, and imaged with BioRad GelDoc XR+ using *Image Lab* v. 5.1 software.

5.8 Sequencing

The correct-looking plasmids were sequenced as Full Service Capillary sequencing at the Finnish Institute of Molecular Medicine (FIMM), Meilahti, Helsinki. In the sequencing reaction samples 150-300 ng plasmid DNA was mixed with 1,6 µl 5µM primers in final volume of 6,6 µl reactions. The sequencing data was analyzed with *Geneious* software (Kearse *et al.*, 2012).

5.9 Cultivation

The cultivations were made in EnPresso B medium following the commercial protocol for multiwell plates. The fresh transformants were inoculated in 2 ml precultures of LB medium supplemented with appropriate antibiotics and incubated in 37 °C, 230 rpm for 6-8h. The used antibiotic concentrations were 50 µg/ml kanamycin and 50 µg/ml chloramphenicol. The EnPresso culture medium was prepared by dissolving one white bag in 100 ml sterile MQ-water and appropriate antibiotics and 25 µl Reagent A was added. The medium was aliquoted to the microplate wells, and the 3 ml well cultures were inoculated with 1:25 of the preculture. The plates were closed with porous membrane and incubated overnight for 15-18 h in 30 °C, 230 rpm. The cultures were boosted and induced with 10x boosting solution with 25 µl Reagent A per 100 ml culture volume and IPTG of 0,5 mM in final concentration. The incubation was continued in 30 °C, 230 rpm for another 24 hours.

5.10 Cell lysis and harvest

The plate cultivations were spun down in Eppendorf centrifuge, 3200 x g for 10 minutes in 4 °C. The supernatants were removed and the cells were resuspended in 1 ml of lysis buffer (20 mM NaH_2PO_4 buffer, pH 7,8, 50 mM NaCl, 1 mg/ml lysozyme, 10 µg/ml DNase I, 10 µg/ml $MgCl_2$, 1 tablet of protease inhibitor cocktail (Sigma-Aldrich S8830) in total volume 150 ml). The plates were incubated on a rocking shaker in 4 °C for few hours or as long as the viscosity of the solutions had increased substantially, and possibly additionally frozen in -20 °C and thawed on icy water. The wells were sonicated one at a time (20 % amplitude, 1 s on, 1 s off) for 20 seconds. The plates were centrifuged in Eppendorf centrifuge on 4500 x g for 45 minutes.

5.11 SDS-PAGE

The SDS-PAGE resolving gel contained 10 % acrylamide (Bio-Rad #1610146), 1,5M Tris-buffer (pH 8,8), 0,1 % sodium dodecyl sulfate, 0,05 % ammonium persulfate and 0,05 % TEMED (Bio-Rad #1610800) and the stacking gel contained 4 % acrylamide, 0,5M Tris-buffer (pH 6,8), 0,1 % sodium dodecyl sulfate, 0,05 % ammonium persulfate and 0,1 % TEMED. The resolving gel mixture (4-5 ml) was pipetted between the glass plates and 0,5-1 ml isopropanol was added on top of the gel to smooth the top surface. After the gel had solidified (after about 45 minutes), the isopropanol was poured off and the stacking gel mixture was added and the combs were set. The gels were ready to use in another 45 minutes.

15 µl of samples were boiled 10 min with 5 µl 4x SDS-PAGE sample buffer (4x solution contains 200 mM Tris-HCl pH 6,8, 40 % v/v glycerol, 8% SDS, 200 mM DTT and bromophenol blue). 10 µl of the mixture was pipetted on gel.

The gels were run in SDS-PAGE running buffer (25mM Tris-base, 0,2M glycine, 0,1% SDS) with 80-130 V, until the blue line of the sample buffer had reached the bottom of the gel. The gels were stained with Coomassie Brilliant Blue R-250

staining solution for either 30 min in room temperature, or 15 min slightly heated up in a microwave. The gels were then destained in destaining solution (20% ethanol, 5% acetic acid) overnight or in destaining solution for 45 minutes and stored temporarily in water. The gels were imaged with BioRad GelDoc XR+ using *Image Lab* v. 5.1 software.

5.12 Alkaline phosphatase assay

Alkaline phosphatase assay (AP-assay) was used to measure the concentrations of the protein of interest in the unpurified samples. For the direct assay from cell suspension 50 µl of the sample was taken from a cooled multiwell plate and mixed with 950 µl Milli-Q water. For the assay of cell lysate or pure protein appropriate dilutions were made. In some measurements the solution was buffered with 50 mM sodium acetate, pH 5, supplemented with 10 mg/ml BSA (Sigma-Aldrich A2153). Samples of 50 µl were pipetted on a microtiter plate, in duplicates or triplicates. 50 µl of liquid *para*-nitrophenyl phosphate substrate (Sigma-Aldrich N7653) was quickly added to all wells and the microtiter plate was immediately put in BioTek Eon/SynergyH1/ Cytation3 microplate reader (Figure 8), which measured the absorbance at 405 nm every minute for 60 minutes (for suspensions), or every 30 seconds for 5-15 minutes (clear samples), or as long as the pattern of the color formation was clear. Also OD_{600} was measured at the starting point for the cell suspension samples. The *Gen5* v. 2.09 software automatically determined the maximal rate of yellow color formation (V_0) as mAU/min within 4 time points in the beginning of the assay. The alkaline phosphatase activity of the protein measured as V_0 corresponds to the protein concentration of the AP-CBM fusion protein.

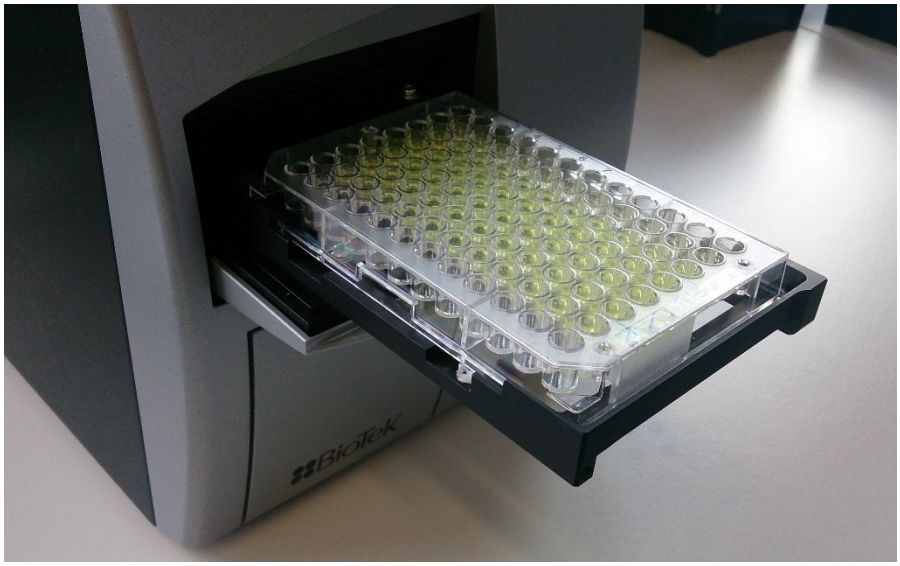

Figure 8. The AP assay was analyzed on a microplate reader.

5.13 Cellulose and chitin binding assay

The proteins were tested with three different substrates, nanofibrillated cellulose (NFC, 6-pass from birch), bacterial cellulose (BC, obtained from Pezhman Mohammadi, grown from *Acetobacter xylinum*, 6-pass) and chitin nanocrystals (ChNC, obtained from Maryam Borghei, made with HCl hydrolysis from commercial chitin, acetylation degree 0,9). In 200 µl total volume, 40 or 80 µg of the substrate was mixed in a tube containing 0,3 µM protein, 10 mg/ml BSA (Sigma-Aldrich A2153), buffered in 50 mM sodium acetate, pH 5. The tubes were mixed well and incubated in room temperature for 60 minutes. The tubes were then centrifuged for 1 min, full speed, and the supernatant was analyzed with SDS-PAGE or alkaline phosphatase assay, in duplicates. Negative control contained equal amount of protein and buffer, and Milli-Q water instead of the substrate. The loss of alkaline phosphatase activity between the control and sample was interpreted to represent the part of the protein bound to the substrate.

6 Results and discussion

6.1 The effect of CyDisCo

All the seven AP-CBM constructs were built in pBR1a expression plasmid and verified by sequencing (ready plasmid of AP-Cel7Aopt obtained from Bart Rooijakkers). The constructs were transformed in BL21(DE3) and BL21(DE3) + CyDisCo by electroporation. In the case of BL21(DE3) + CyDisCo, the CyDisCo plasmid had been earlier transformed in BL21(DE3) and electrocompetent cells were made of this strain (electrocompetent cells obtained from Bart Rooijakkers).

The experiments were started by comparing the constructs in BL21(DE3) strain. The cultivated and induced cells of Cel7A construct in BL21(DE3) strain with and without CyDisCo plasmid were lysed and the supernatant was used for the AP-assay and SDS-PAGE (Figure 9). First of all the production levels seemed to be greatly enhanced by the CyDisCo plasmid even though those cells were grown in two antibiotics. The AP-assay values for V_0 seem to correlate quite well with the SDS-PAGE results, so the functionality of the AP-assay was verified. The presence of the AP-CBM protein was seen in the SDS-PAGE image as 55,4 kDa protein. However, the CyDisCo proteins Erv1p (21,6 kDa) and PDI (55,4 kDa) were also seen on the gel image, and especially the PDI caused problems in interpreting the results as PDI is almost exactly of the same size as AP-CBM proteins.

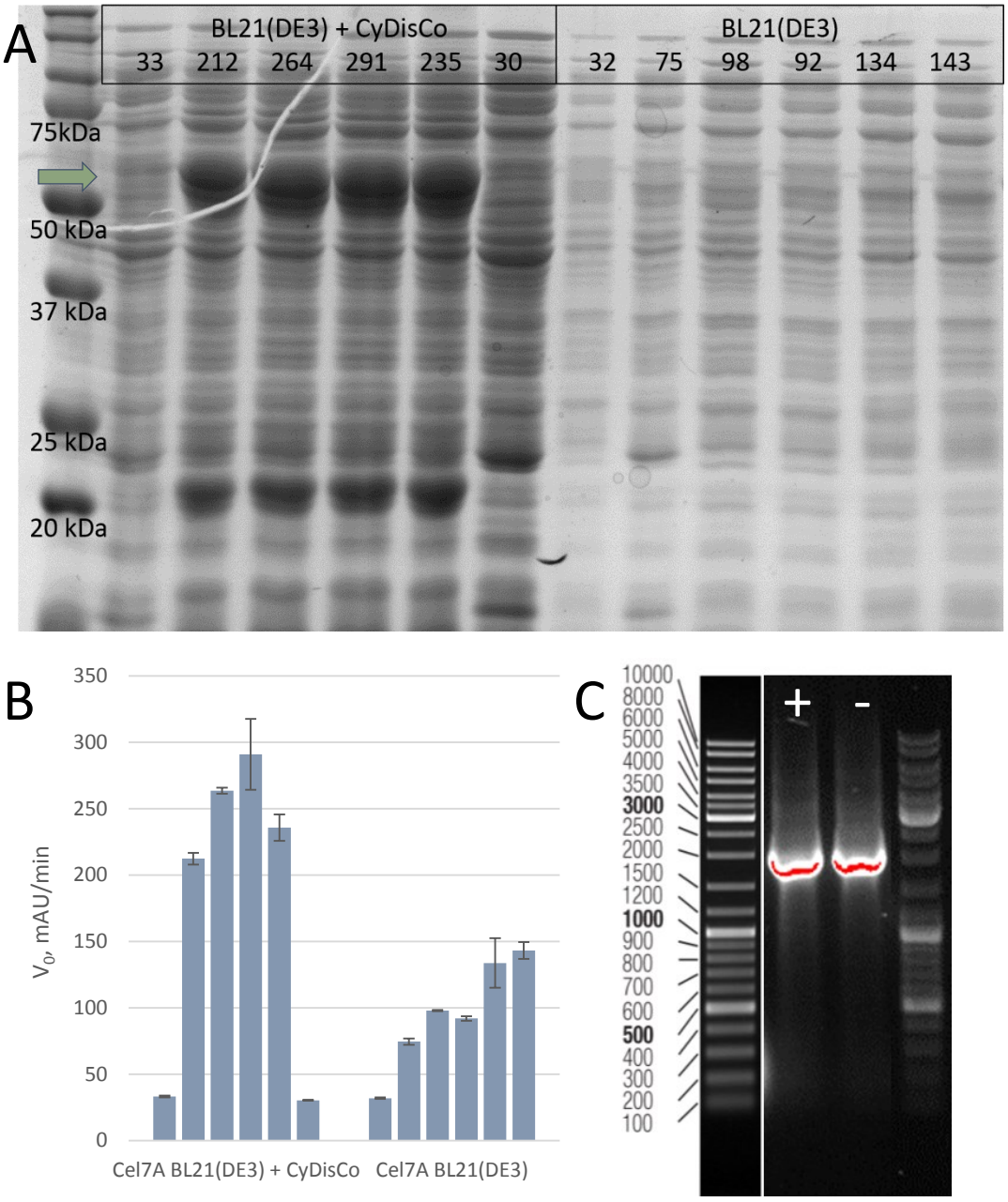

Figure 9. The A) SDS-PAGE and B) AP-assay results from the cell lysis supernatant of Cel7Aopt clones in BL21(DE3) strains with and without CyDisCo. The V_0 values (marked also on the SDS-PAGE image) indicate the initial activity of the AP enzyme (55,4 kDa) present. The error bars represent standard deviation between replicates. C) The presence of the AP-CBM gene was verified with colony PCR from expressing and non-expressing clone from BL21(DE3) + CyDisCo strain.

Another remarkable observation was that the variation between the clones turned out to be massive. The previous results of several hundredfold differences in the

expression of different constructs (Ikonen, unpublished data) have more likely been due to a poor selection of the clone, and not due to the variation caused by different protein or nucleotide sequences of the CBM. Additionally, in the cases where the AP-activity was low neither AP protein nor either of the CyDisCo proteins were noticed on SDS-PAGE. However, according to colony PRC made to amplify the AP-CBM gene, the 1949 bp coding sequence was found in both expressing and non-expressing clones (Figure 9). There is also a possibility that something else had been broken in the expression system, like the T7 RNA polymerase gene, which was not tested with colony PCR. However, this did not explain the disappearance of CyDisCo proteins as they were regulated by *tac* promoter, which is directly inducible by IPTG (de Boer *et al.*, 1983) and not dependent on the T7 RNA polymerase.

6.2 Testing the AP-assay method

As the AP-assay proved to be a good way to measure the AP-CBM protein concentration in a lysate which additionally contained all other proteins and metabolites from the cells, the method was further developed. Cell lysis is a time-consuming step, especially with microwell plates where every well needs to be resuspended to lysis buffer and sonicated separately. The AP-assay was tested from the cell suspension, the clear growth medium without cells, and the cell lysate of transformants in BL21(DE3) strain (Figure 10).

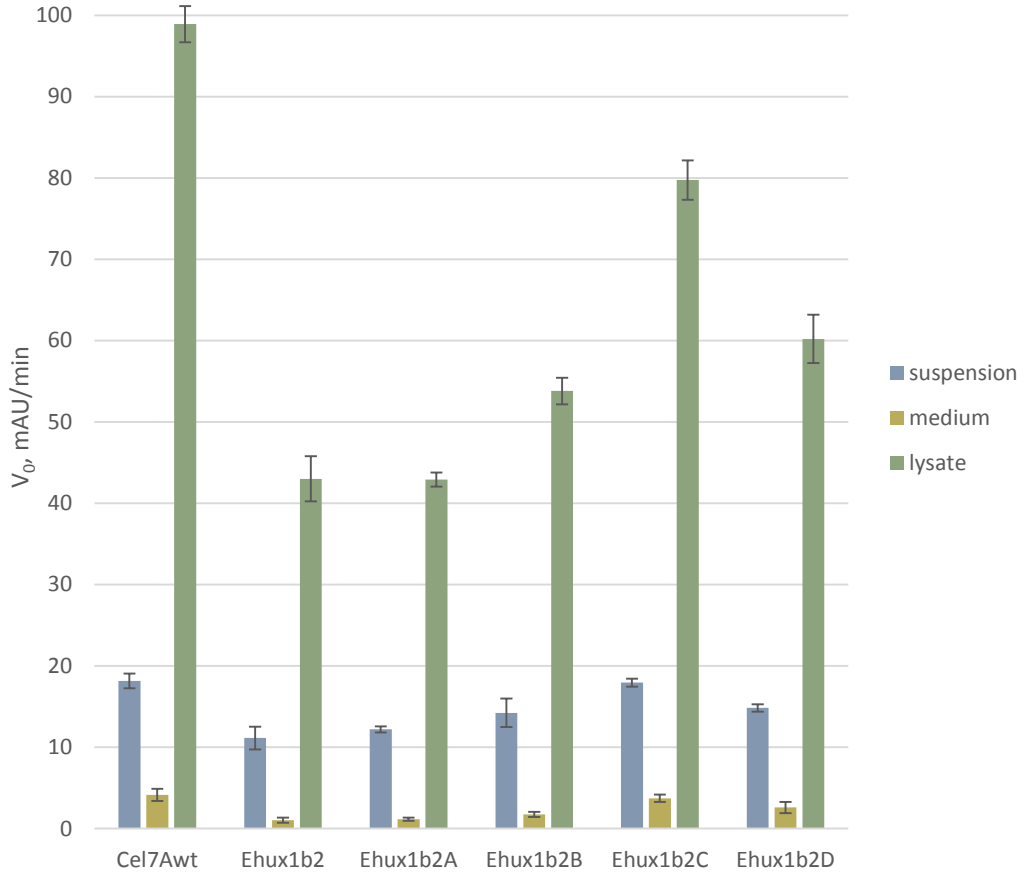

Figure 10. The comparison of AP-assay results using cell suspension, growth medium without cells of the cell lysate from BL21(DE3) strain as sample type.

The same order in the AP-activities was maintained when measuring from different sample types. The growth medium would have been an ideal sample type as it contained no particles and was more stable compared to cell suspension. On the other hand, the brown-yellow color of LB medium could affect the absorbance readings. However, since the AP-activity was so weak in the case of growth medium, and the error defined by standard deviation was relatively big, the cell suspension was chosen as the sample type for future experiments. The challenges in using cell suspension are that the cells might sediment at the bottom of the microtiter plate preventing absorbance readings, so the plate needs to be shaken properly, and that the possible growth of the cells may bias the color formation rates. To prevent the sedimentation and optical issues due to particles and existing

yellow color, the cell suspensions were always diluted 1:20 to water before taking the sample for the assay.

The V_0 values were defined by the maximal rate of increase in yellow colour at 405 nm within 4 time points during the first 4 minutes. The activity was determined in such early phase of the assay to get the initial value for the color formation rate, but this method also caused problems as there were false positives. When comparing the AP-assay results to the absorbance increase curves and SDS-PAGE images, the false positives could be revealed (Figure 11). As little difference as from 11,6 mAU/min to 10,1 mAU/min was found to be a difference between expressing and non-expressing clone in the V_0 values. The color formation curves of the false positives had distinguishable shape whereas the expressing samples had a curve that was smooth and linear or slightly decelerating by time. Consequently the AP-activity values up to 10 mAU/min, measured from cell suspension diluted 1:20, needed further attention. Fortunately, when CyDisCo was not used, nearly all clones expressed the protein, so there was no such risk for detecting false positives.

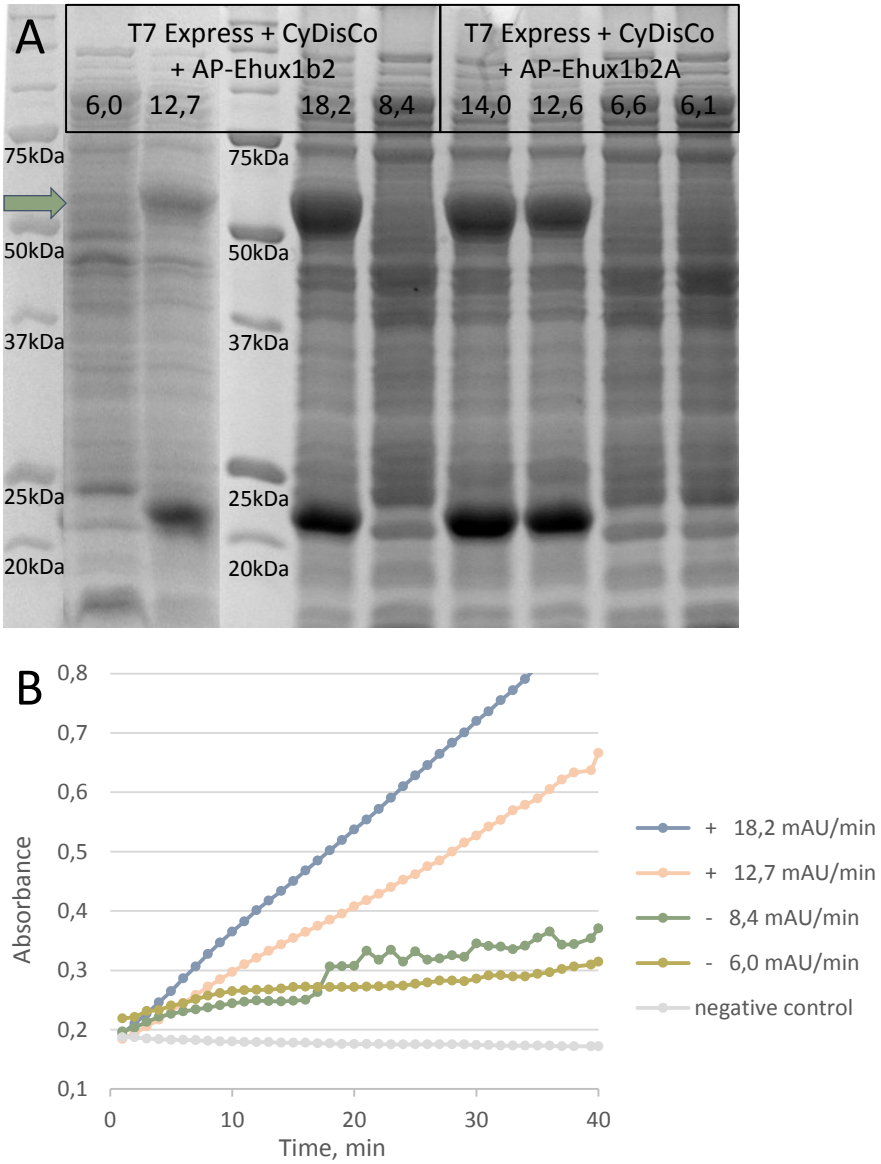

Figure 11. A) The AP-activity is not unequivocal measure of the protein expression. The AP-activities, V_0, of each sample of Ehux1b2 and Ehux1b2A are indicated on the top row (mAU/min). B) AP-assay activity curves of the corresponding Ehux1b2 clones represent an example of false positive samples. Two samples (+) were expressing well the AP-Ehux1b2 protein and two others (-) were false positives in V_0 values.

6.3 The correlation of the AP-activity and protein concentration

The correlation between AP-activity and protein concentration was determined by analyzing a dilution series of Cel7Aopt protein sample of known concentration

(made by Bart Rooijakkers) with the AP-assay (Figure 12). At first the correlation was studied by diluting the 100 µM protein in H_2O. Upon addition of the AP substrate *para*-nitrophenyl phosphate (*p*NPP) to samples over 2 µM of protein, the solution turned yellow too fast. With such concentrated samples the technique to add substrate to all wells with a multipipette and then inserting the microtiter plate in the microplate reader is not fast enough to be able to detect the initial color formation rates, as the rates are already decreasing by the time of reading the absorbances. On the other hand, concentrations under 40 nM were suspicious as they give activity rates which were close to the values of false positives. Consequently, the range for AP-activity measurements for protein concentration determination was found to be between 0,2-2 µM.

Still making a linear standard curve was difficult and the error in pipetting such small volumes and its accumulation in the more dilute samples may have accounted for a considerable proportion of the fluctuation. When having such dilute protein samples the unspecific binding of the protein to the tube walls and pipette tips was most likely a significant factor to the erroneous pipetting. Nevertheless, bovine serum albumin (BSA) could be used to reduce the unspecific binding as the BSA tends to bind to the possible targets more strongly (Linder, 1996). Additionally the protein solution was buffered to pH 5 with a 50 mM sodium acetate buffer, because this way the pH of the lysates could be standardized. The addition of 10 mg/ml BSA turned out to help the linearity of the standard curve tremendously (Figure 12, Equation 1). Surprisingly also the used BSA itself seems to have some AP activity as the negative control without AP-CBM has AP activity of 2,7 mAU/min. Although BSA itself should not exhibit alkaline phosphatase activity, the method of purification used at the manufacturing, purification by ethanol fractionation, may not be precise enough to eliminate all other protein from the BSA.

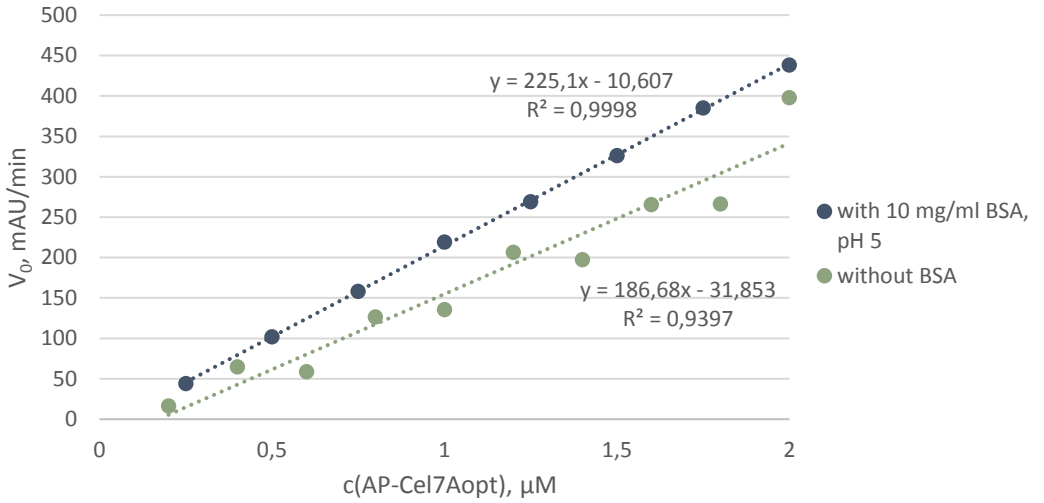

Figure 12. The determination of the correlation of protein concentration and AP-assay and the effect of BSA in the accuracy of the measurements.

$$c(protein, \mu M) = 0{,}0044 * V_0 + 0{,}0471 \quad (1)$$

6.4 Clone variation in BL21(DE3) and T7 Express strains

The clone variation was further studied with all 7 constructs in BL21(DE3) and T7 Express strain without CyDisCo. To get extensive data on the clone variation, 24 (BL21(DE3)) or 10 (T7 Express) colonies of different size per construct were selected to the cultivation and direct AP-assay analysis from the growth suspension (Figure 13). Additionally some of the samples were run on SDS-PAGE to verify the results of AP-assay.

Unlike in the case of CyDisCo, where some clones did not express any recombinant protein, surprisingly the plain BL21(DE3) strain was quite stable in the expression and no significant differences were observed between clones. T7 Express had also quite little variation, but the poor expression of a single clone caused significant standard deviation in Ehux1b2D. As the plain strains did not show the variation previously seen with CyDisCo, it seemed that the CyDisCo had an unknown effect to the strains which caused some clones to express well but inhibited all expression from others. Nevertheless, CyDisCo proteins are unlikely to cause

major changes in the cell metabolism as they only assist in the folding. Even so, the colony picking was an important phase and several clones should be always tested in small scale before choosing one for larger scale production.

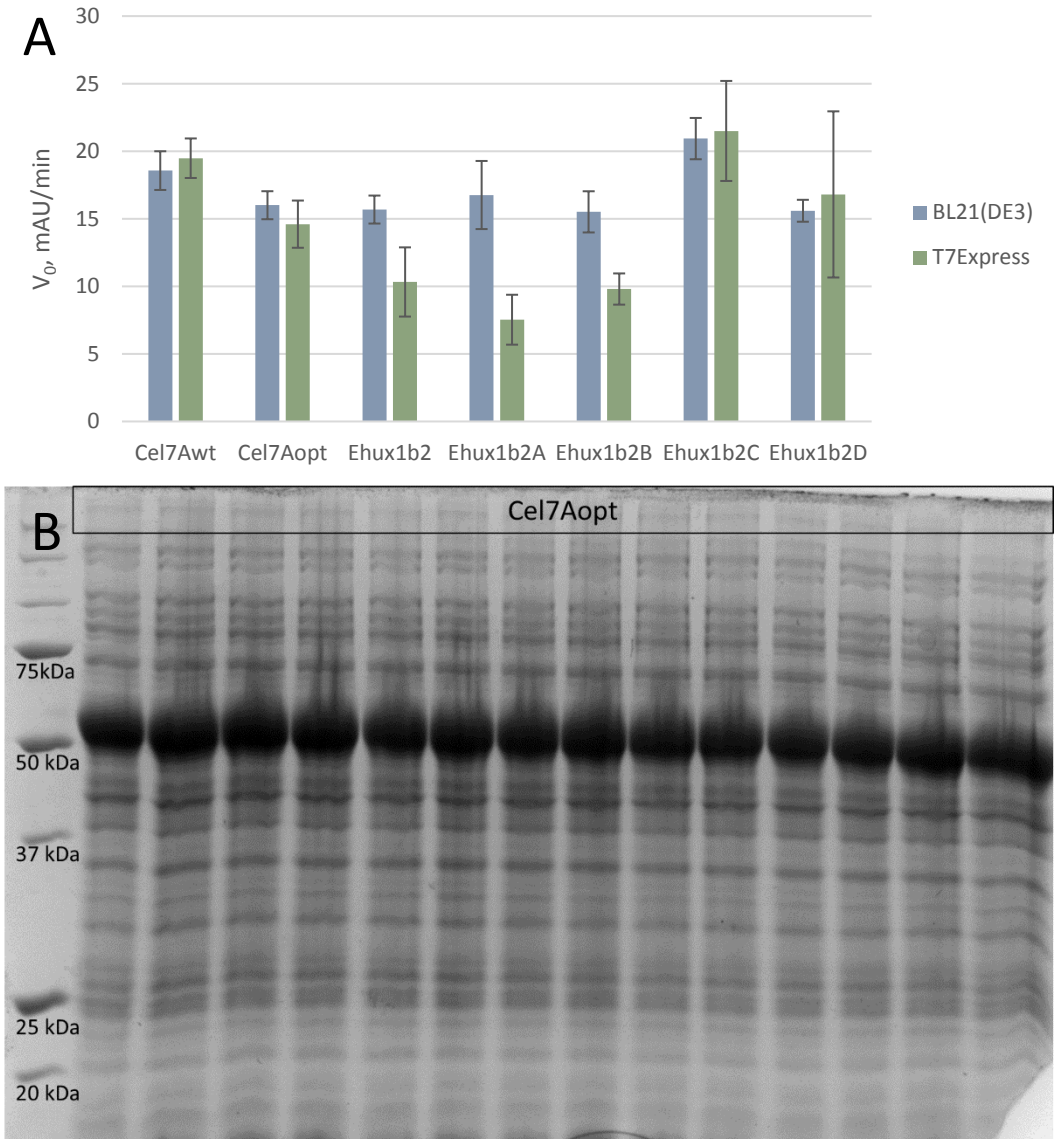

Figure 13. The clone variation in different constructs in BL21(DE3) represented by A) variation of 24 or 10 clones per construct in direct AP-assay from growth suspension of BL21(DE3) and T7 Express, respectively, and B) variation within 14 Cel7Aopt construct from BL21(DE3) on SDS-PAGE.

The expression of different constructs was not completely consistent in both strains, which indicates that the strains have some preferences for the sequences.

Nevertheless, the expression of Ehux1b2 was clearly enhanced by the substitution of the C-block from Cel7Aopt (Ehux1b2C). Additionally the wild-type sequence for Cel7A turned out to be expressed better than the optimized sequence. The difference between these synonymous sequences is not big and it is a bit more substantial in T7 Express. However, this proves the fact discussed in the earlier chapters that the optimization is not straightforward and the organisms have preferences that are not generally known.

6.5 Transformation problems with CyDisCo

As BL21(DE3) showed little clone variation between clones, but significant variation had been spotted with CyDisCo plasmid, a similar comprehensive study was tried to make. However, there were constant problems in making a CyDisCo containing strain, which would express. The clones transformed to the original batch of BL21(DE3) + CyDisCo electrocompetent cells (obtained from Bart Rooijakkers) used in the first experiments expressed 60% (Figure 9, Figure 14). Nevertheless, following the same protocols and remaking competent cells from the same strain resulted in no expression of any construct. For some unknown reason, none of the later made batches of any kind of competent cells in either of the *E. coli* strains containing CyDisCo plasmid were expressing any of the recombinant proteins (Table 1). Finally, the double transformation, where both plasmids were transformed at the same time was found to work.

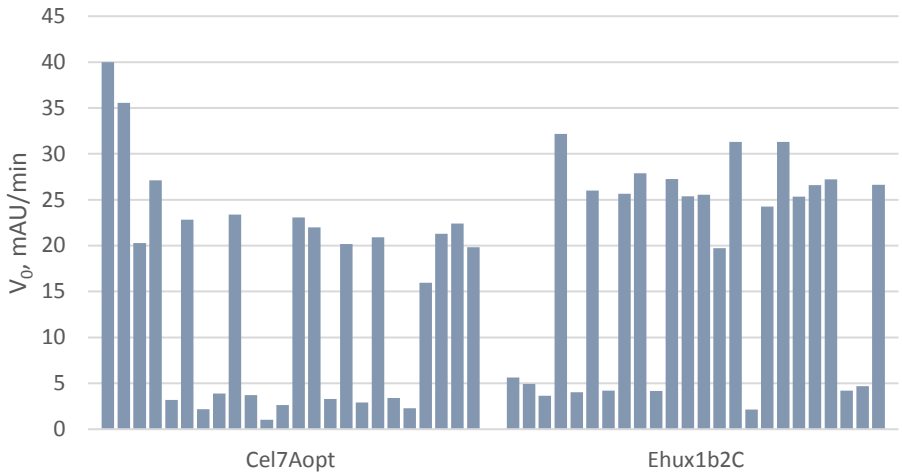

Figure 14. The expression of clones made by electroporation of AP-CBM -plasmid to BL21(DE3) + CyDisCo electrocompetent cells of the first batch.

Table 1. The success of different strategies for CyDisCo strain transformations.

Competent cell type	**Plasmids to be transformed**	**Transformation efficiency**	**Expression**
BL21(DE3) + CyDisCo chemically competent	AP-CBM	+++	-
BL21(DE3) + CyDisCo electrocompetent	AP-CBM	+	+/- One batch expressed 60%
T7 Express + CyDisCo chemically competent	AP-CBM	+++	-
T7 Express + CyDisCo electrocompetent	AP-CBM	++	-
T7 Express chemically competent	CyDisCo AP-CBM	++	+ Expressed 40%

Apparently the addition of CyDisCo plasmid in the strain caused stress to the cell, which supposedly prevented the expression of all recombinant proteins. The competent cells of the CyDisCo containing strains were also tested for expression without the AP-CBM plasmid and unexpectedly even the CyDisCo proteins were not expressed (Figure 15). It is notable, that in the plasmids that were used in this thesis, the LacI repressor was positioned in the AP-CBM plasmid and not in CyDisCo. There is also available another version of CyDisCo plasmid which would

have the repressor, but it was not tested in this study. However, as the CyDisCo was transformed first, its genes may not have been repressed enough, as the tac promoter was not repressed in the absence of LacI. Usually the leaking is most harmful when the expressed proteins are toxic to the cells, which the Erv1 and PDI should not be. In the case the cell regards the new DNA as detrimental, there might be genetic rearrangements, which may splice out some elements. (L. Ruddock, personal communication, 18.5.2016)

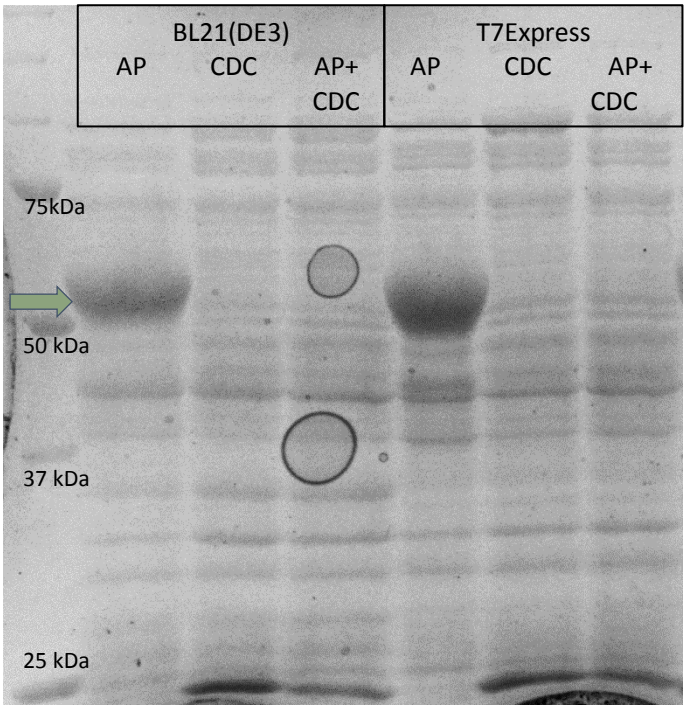

Figure 15. The BL21(DE3) and T7 Express strains containing CyDisCo (CDC) did not express even without the AP-CBM (AP) plasmid.

In the case of the first batch of BL21(DE3) + CyDisCo competent cells, there must have been less stress for the cells as everything still worked in 60% of the clones. The double transformation is one way to prevent the leakage as the repressor is introduced simultaneously with the CyDisCo. Nevertheless, there was still only 40 % of the clones which expressed, so the quality was still not good. Transforming first the AP-CBM plasmid and then the CyDisCo would be worth trying to find out this mystery.

6.6 Comparison of constructs in T7 Express + CyDisCo transformation

As the double transformation solved the issue with CyDisCo, a proper expression test was made in T7 Express strain. Ten clones of each construct were analyzed with AP-assay (Figure 16) and SDS-PAGE (Figure 17A). The SDS-PAGE results showed qualitatively the same results as AP-assay but quantitatively the slight differences were hard to recognize. On the other hand, there were many false positives. When comparing the SDS-PAGE samples to the AP-activities of the corresponding samples, the false positives could be ruled out.

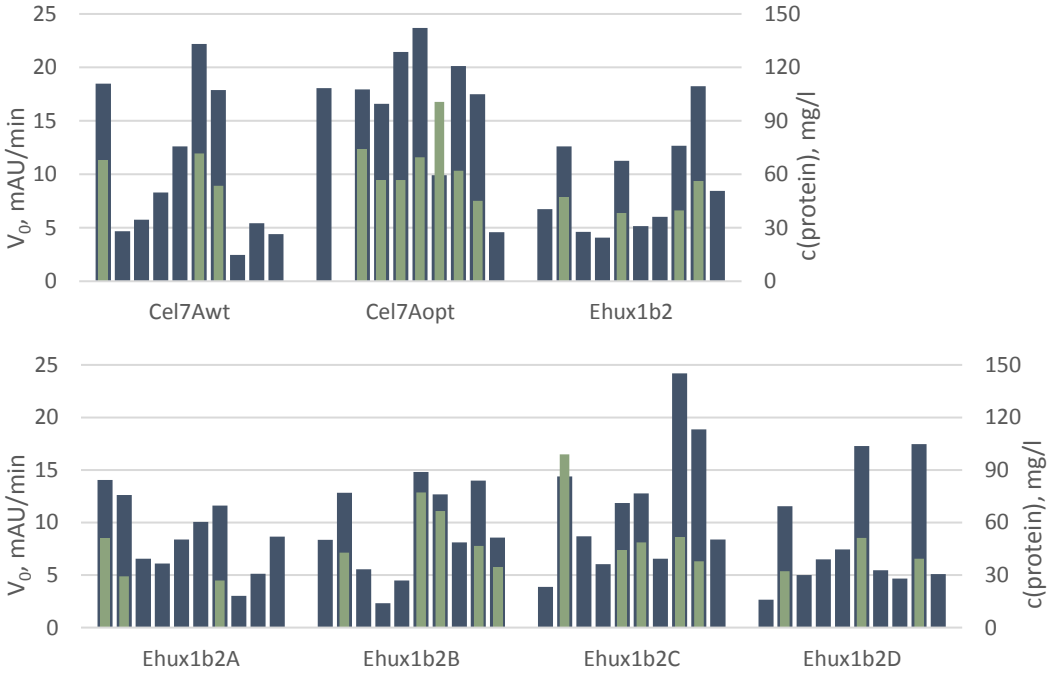

Figure 16. The variation in the expression of different constructs in T7 Express + CyDisCo made with double transformation (blue, left axis) and the corresponding protein yields in the cultivations analyzed from the lysates (green, right axis).

Based on the results, the expressing clones were selected to be lysed and their proteins were harvested. The clear lysates were then analyzed with AP-assay in 50mM sodium acetate buffer, pH 5, containing 10 mg/ml BSA. Using the standard curve made with purified protein (Figure 12), the concentrations in each sample were determined and proportioned to the culture volume. At best, the clones of

Cel7Aopt and Ehux1b2C reached the expression of 100mg/l, which is sufficient protein yield for this purpose. However, the expression of other *E. huxleyi* constructs fell behind, the concentrations being mostly within 30-60 mg/l.

When comparing the results from lysates to the results from AP-assay from cell suspension or SDS-PAGE gel images, the results within a sample were not always consistent. The pipetting was a major source of error in all steps. In the cell suspension the mixing of the suspension may have been inadequate and resulted in more or less dense solution at the dilution. The protein concentration of the lysates were affected by the success of cell lysis. The lysis itself was also a tricky process, as the cells needed to be sonicated enough to extract all the protein, but too intensive sonication would have caused the protein to degrade. The small volumes also made the relative error rate bigger. The remaining supernatant from the growth medium removal as well as the collecting of the lysate after cell lysis and centrifugation were steps that affected the volume and thus the concentration. Additionally, in the methods used, only the soluble fraction of the protein was taken into account. Although that fraction is what we were aiming for, a significant proportion of the protein may have been in insoluble form.

When harvesting the proteins, an interesting observation was made as after removing the growth medium, the cell pellets were coloured differently – the cells that expressed the recombinant proteins formed more yellow cell pellet (Figure 17). As the proteins expressed were not coloured this was likely due to a change in the metabolism, which caused accumulation of a yellow-coloured compound. Although the difference in the colour was quite vague, it could still be used as preliminary screening of clones to avoid the wasting of reagents to test the grey clones.

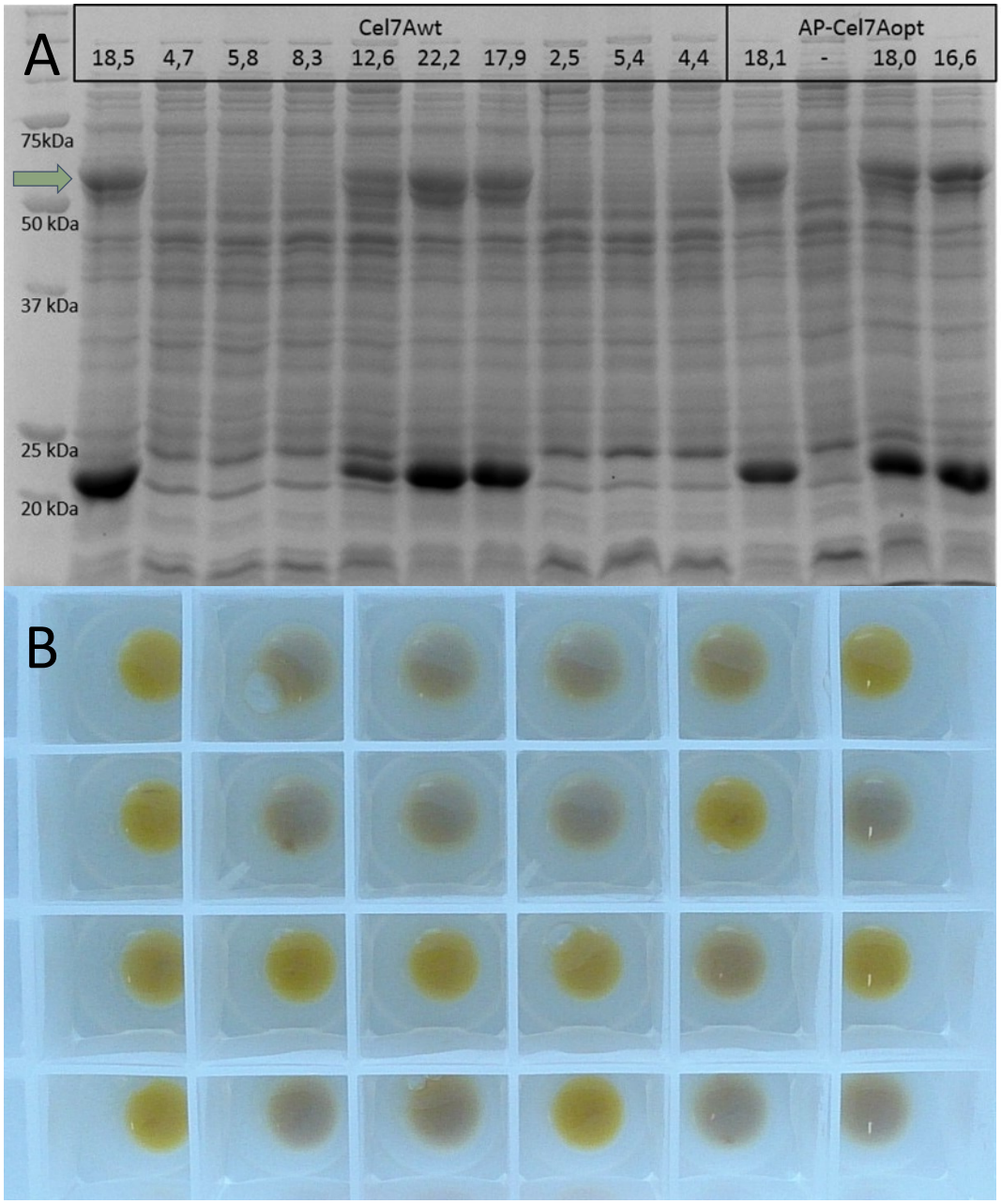

Figure 17. A) The SDS-PAGE image of the first clones of first CyDisCo + AP-CBM double transformations in T7 Express. B) The expression could be roughly identified by the color of the cell pellet after removal of the growth medium. The order on the SDS-PAGE gel was the same as reading order in the cultivation on 24-well plate.

6.7 Comparison of different strains

To compare the expression between strains, two separate experiments were performed. One contained specific information about the expression in T7 Express and two constructs were additionally cultivated in BL21(DE3) (Figure 18A). To further compare the different strains and also investigate the additional value of CyDisCo to T7 Express, which was expected to be a better strain than BL21(DE3), three constructs from the different edges of the production range where chosen to the reduced amount of variables. They were recultivated simultaneously to eliminate the differences in the growth conditions. As in the expression with CyDisCo some of the clones did not express at all, and the good clone would need to be selected anyway by expression tests before larger scale production, the mean values of AP activity of only the expressing clones were compared (Figure 18B).

In the first dataset T7 Express seemed to be significantly better strain for the production. The optical density of the sample gave some explanation to the differences as T7 Express was constantly higher in density. The optical densities measured on the microtiter plate from the 1:20 diluted sample, further diluted by 1:2 upon addition of the substrate, most likely contained inaccuracy from the pipetting of the cell suspension, and thus the calculation of the actual cultivation OD was not sensible. Nevertheless, the density should have accounted for most of the variation between the strains. The results of the second dataset were ambiguous when compared to the first experiment. All the V_0 values for T7 Express were almost cut to half. So were the densities, but the ratio of V_0 and OD_{600} was still smaller. Additionally the superiority of the two plain strains was switched in the second experiment, albeit the difference was small. In general, the density of T7 Express cultures was higher. The additional value of CyDisCo to the protein production was rather small, but it did affect negatively to the growth of the cells, most likely due to the cultivation in two antibiotics and the increased stress caused by maintenance of two plasmids. However, the usage of CyDisCo should affect

positively in the folding of the AP-CBM protein and consequently the proteins expressed with CyDisCo should be more functional. A clear conclusion of the different strains is that the C-block increased the expression of Ehux1b2 to the level of Cel7A in all of the strains.

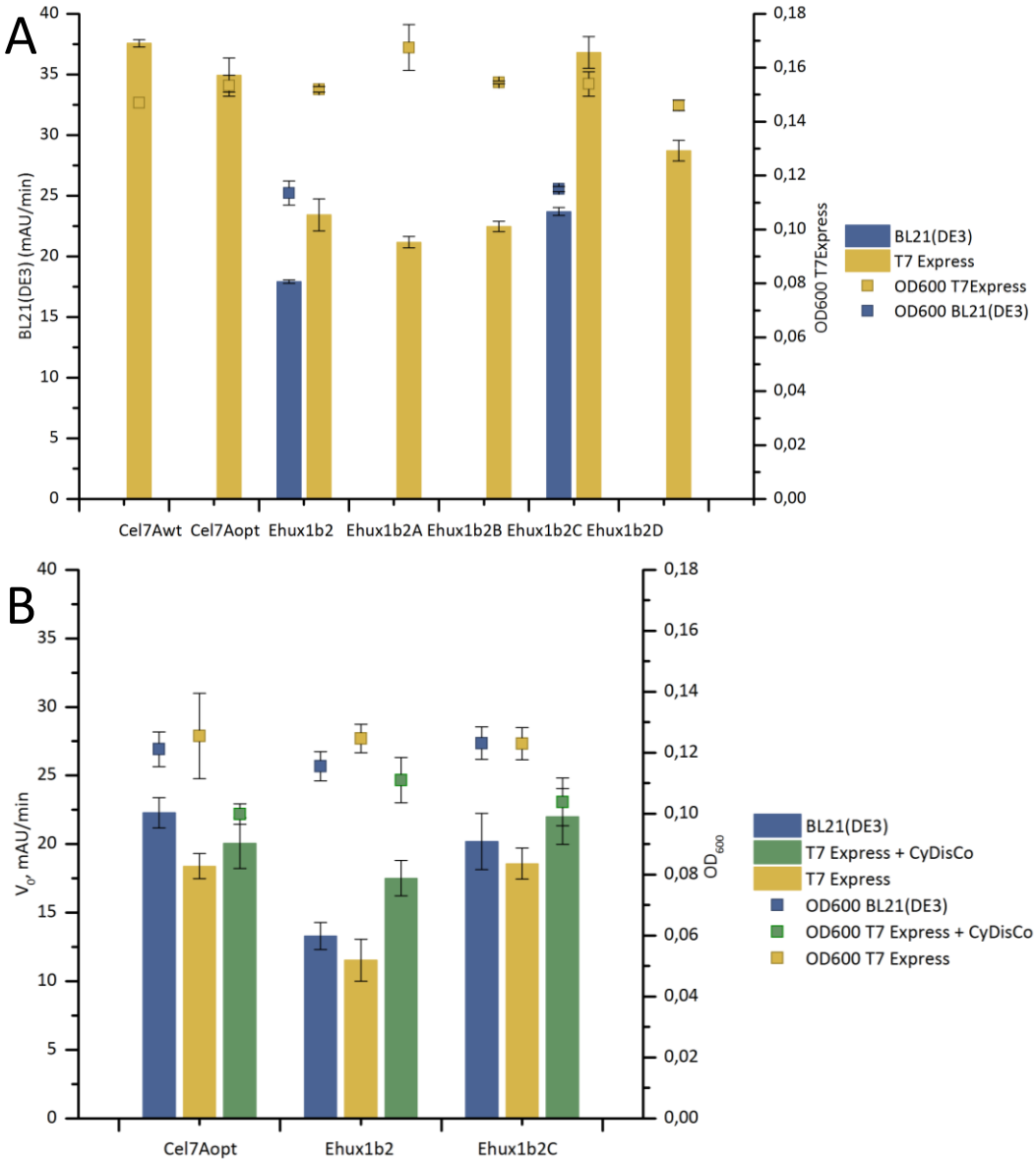

Figure 18. The comparison of AP-activities and optical densities of A) T7 Express with BL21(DE3) with two constructs and B) three constructs in three strains.

6.8 Binding affinities of the proteins

The binding properties of each CBM were tested by simple binding assays with several substrates. In addition to the expected substrate, nanofibrillated cellulose (NFC), the binding affinities towards bacterial cellulose (BC) as well as chitin nanocrystals (ChNC) were tested. By comparing the supernatants of samples with and without substrate, the percentages of the protein bound were determined (Figure 19).

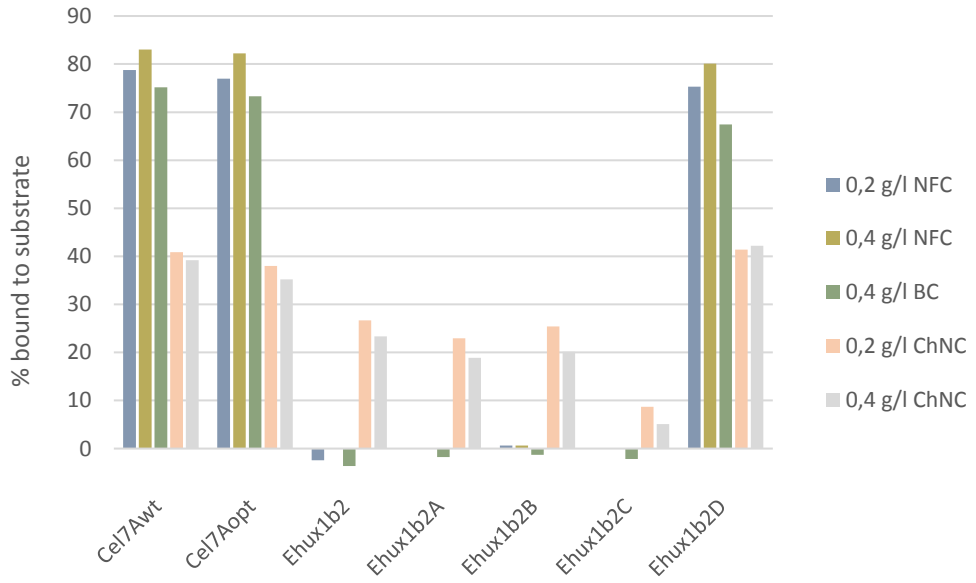

Figure 19. The binding affinities of each protein (0,3 µM) towards nanofibrillated cellulose (NFC), bacterial cellulose (BC) and chitin nanocrystals (ChNC), displayed as percentage of bound protein.

As the alkaline phosphatase was the same in all constructs, the differences in the AP-assay activities resulted from the different amounts of protein instead of different functionalities of the proteins. In other words, the binding affinity towards nanocellulose was highly dependent on the CBM sequence. As expected, the reference protein Cel7A bound well to NFC and slightly less to BC. It was also clearly seen that chitin was not its primary substrate. Additionally, the substrate concentrations did not have significant effect on the results, which indicates that there was excess of the substrate and there was no competition in the binding.

Surprisingly, despite the similarity of the sequence to Cel7A, the coccolithophorid sequence Ehux1b2 did not bind at all to either type of cellulose, and the chitin affinity was slightly smaller.

However, one hybrid sequence, Ehux1b2D, showed significantly enhanced properties in binding compared to Ehux1b2 and other hybrids, which barely bound at all. With just five amino acid substitutions in the C-terminus, the binding affinity of Ehux1b2 could be restored to the level of Cel7A. The D-segment (Figure 20) contained the two adjacent tyrosines, which are found to be important residues in the binding to cellulose, and the asparagine with a less clear contribution to the binding (Linder et al., 1995). These results confirm the importance of the specific residues in the C-terminal end of the CBM sequence. Additionally, as the function is recovered by just few substitutions, this result confirms that the Ehux1b2 sequence folds the same way as Cel7A so the proteins are structurally related.

Figure 20. The change of only 5 amino acids in the sequence of Ehux1b2 restored the binding ability. The two tyrosines supposedly play the biggest role in the cellulose binding (Image by *Geneious*, Kearse *et al.*, 2012).

As the coccolithophorid protein Ehux1b2 seemed not to bind either in cellulose or chitin, the binding affinities of the coccolithophorid CBMs towards other polysaccharides would be interesting to test. Reportedly unknown type of polysaccharides are contributing to the calcification process of *E. huxleyi* (Kayano

et al., 2011). The full binding isotherms would also tell more details about the binding, but making a binding isotherm would require more samples and consequently more protein, which could not be achieved with the cultures on microwell plates but require bigger culture volumes.

6.9 Sequence analysis of the proteins

Inspired by the significant contribution of D-block to the binding, CBM1 sequences were aligned to find out naturally existing mutations within cellulose binding enzymes. Amongst the nearly thousand CBM1 sequences available, 27 sequences were selected by random, based on their usage of cellulose, xylan or chitin as a substrate, and the ones with low sequence similarity were excluded. The sequences were aligned to Cel7A sequence (Figure 21, GenBank accession numbers in Appendix 2).

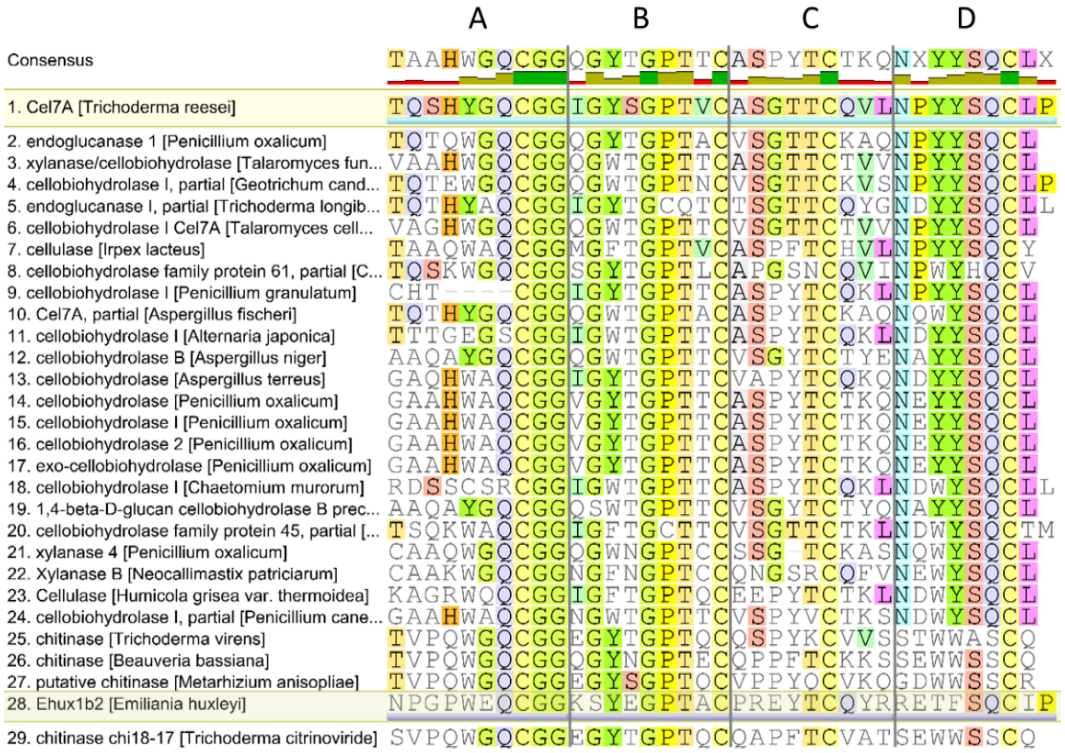

Figure 21. Multiple sequence alignment of 27 CBM1 sequences selected by the similarity of structure to Cel7A and cellulose-related activity (Image by *Geneious*, Kearse *et al.*, 2012).

The comparison revealed that the residues of D-block are highly conserved among cellulose binding CBM1 sequences, so that region must have an important contribution to the protein function. Similar conclusion could not be made from other segments, as the residues that were not the same in Ehux1b2 and Cel7A, were widely varying among different CBM1 sequences. To conclude, the alignment supports the observation that the addition of D-block could restore the function. In addition to the two tyrosines in D-block, also the asparagine is conserved in most of the selected sequences, and only chitin binding CBMs lack it. Possibly also the proline makes the structure more rigid and modifies the structure to better bind to cellulose. Some chitinases were also included in the multiple sequence alignment, but the similarity of the Ehux1b2 D-block is not significantly better with them either. This result suggests that the main substrate of Ehux1b2 is some other polysaccharide, possibly not even chitin.

In segments A and B the conservation of amino acids is approximately the same within Ehux1b2 and Cel7A as with all other CBM1 sequences. All of the most conserved residues are also present in the sequence of Ehux1b2. Nevertheless, most likely these highly conserved residues form the fold for the protein, as its structure seemed to be functional when the important residues of D-block were substituted.

7 Conclusions

The codon optimization of a gene sequence was found to be troublesome as there is still a lack of experimental evidence how different kinds of sequences work in different hosts. The desirable properties of a gene sequence are defined by several overlapping variables, and the simultaneous optimization of them all is challenging. In the experimental research the sequence optimization was studied by expressing shuffled CBM1 homologues in four *E. coli* strains and testing the binding affinities towards nanosized cellulose and chitin.

The alkaline phosphatase assay was found to be a good qualitative way to detect the protein expression. The AP-assay was used as a way to measure AP-CBM protein concentrations from cloudy cell suspensions as well as from unpurified lysates that contained all proteins from the cells. However, in the measurements with cloudy samples the assay was not robust enough and could result in false positives, which could be revealed on SDS-PAGE. Additionally the detection range was quite narrow, between 0,2-2 µM. The reliability of AP-assay could be enhanced by using clear samples and adding BSA to the mixture to prevent unspecific binding. In this production scale and type of research, the property to detect specifically the protein of interest from a mixture was crucial, as the amount of samples exceeded the number that could have possibly been purified.

In the protein expression the wild type sequence of Cel7A was occasionally performing better than the optimized version, which emphasized the difficulty of optimization. In T7 Express strain containing CyDisCo, Cel7Aopt was expressed 101 mg/l at the best, but the original homologue Ehux1b2 only 38-56 mg/l. The hybrid CBM Ehux1b2C was expressed nearly as well as Cel7A with 99 mg/l, which indicated that the substitution of C-block with the sequence from Cel7A brought favorable properties to the expression. The other block substitutions did not have significant effect in the expression. Although the proteins were shuffled segment-wise, and the codon usage derived from the optimized parental sequence of the segment, the shuffling could still cause both global and local changes in the codon usage and that way it may have affected the expression.

T7 Express and BL21(DE3) were found to be approximately equally good in the production, although the results were varying. According to the genotype, T7 Express should be more stable strain, and in the experiments it grew to a slightly higher density. The CyDisCo system could enhance the protein production up to 100%, but it also brought problems with it. The clone variation without CyDisCo plasmid was almost nonexistent compared to the situation with CyDisCo where only 40-60 % of clones expressed. In order to make at least some clones express,

the CyDisCo plasmid had to be transformed simultaneously with the AP-CBM plasmid to the strain. The order of the expression levels of different construct followed a similar pattern in all strains – both versions of Cel7A at the top alongside Ehux1b2C. The original coccolithophorid sequence Ehux1b2 was expressed the least. The other block substitutions had varying amount of positive contribution to the expression.

The binding abilities of all seven proteins were tested with nanofibrillated cellulose, bacterial cellulose and chitin nanocrystals. About 80 % of Cel7A bound to cellulosic substrates, and 35-40 % bound to ChNC. On the contrary, Ehux1b2 did not bind to cellulose at all, and was also poor at binding to ChNC with 25 % of the protein bound. However, when the D-block of Ehux1b2 was substituted with the corresponding sequence from Cel7A, the binding affinities towards all three substrates were restored. No such behavior was observed with any other block substitutions, and the C-block substitution also diminished the chitin binding affinity.

A comparison with other CBM1 sequences revealed that the D-block contained the most such amino acids that are conserved normally within cellulose binding modules, but not in Ehux1b2 sequence. Therefore they could be regarded as important residues for the binding. In addition, the evidence suggested that Ehux1b2 is not a cellulose binding module and possibly not even chitin binding module, but might have activity towards some other polysaccharide. However, it must be structurally similar to Cel7A as substitution of five amino acids could give it the cellulose affinity.

To conclude, the relative expression levels were not dependent on the *E. coli* strain used, and the reference sequence Cel7A was both produced the best and it also had the highest affinity towards cellulose and chitin. The substitution of the D-block from Cel7A could make the otherwise inactive Ehux1b2 to bind to cellulose.

References

Abitbol, T., Rivkin, A., Cao, Y., Nevo, Y., Abraham, E., Ben-Shalom, T., Lapidot, S., and Shoseyov, O., Nanocellulose, a tiny fiber with huge applications, *Curr. Opin. Biotechnol.* **39** (2016) 76–88.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P., *Molecular Biology of the Cell*, 5th ed., Garland Science, New York 2008, 1392 p.

Allert, M., Cox, J.C., and Hellinga, H.W., Multifactorial Determinants of Protein Expression in Prokaryotic Open Reading Frames, *J. Mol. Biol.* **402** (2010) 905–918.

Angov, E., Hillier, C.J., Kincaid, R.L., and Lyon, J.A., Heterologous protein expression is enhanced by harmonizing the codon usage frequencies of the target gene with those of the expression host, *PLoS One* **3** (2008) 1–10.

Arola, S., *Biochemical modification and functionalization of nanocellulose surface*, Doctoral dissertation, VTT Publications 102, Espoo 2015, 95 p.

Brownlee, C., Wheeler, G., and Taylor, A.R., Coccolithophore biomineralization: New questions, new answers, *Semin. Cell Dev. Biol.* **46** (2015) 1–6.

Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A., and Zhang, F., Multiplex Genome Engineering Using CRISPR/Cas Systems, *Science* **339** (2013) 819–823.

Crameri, A., Raillard, S.A., Bermudez, E., and Stemmer, W.P., DNA shuffling of a family of genes from diverse species accelerates directed evolution, *Nature* **391** (1998) 288–91.

de Boer, H.A., Comstock, L.J., and Vasser, M., The tac promoter: a functional hybrid derived from the trp and lac promoters, *Proc. Natl. Acad. Sci. U. S. A.* **80** (1983) 21–5.

de Marco, A., Strategies for successful recombinant expression of disulfide bond-dependent proteins in Escherichia coli, *Microb. Cell Fact.* **8** (2009) 26.

Engler, C., Gruetzner, R., Kandzia, R., and Marillonnet, S., Golden Gate Shuffling: A One-Pot DNA Shuffling Method Based on Type IIs Restriction Enzymes, *PLoS One* **4** (2009) e5553.

Engler, C., Kandzia, R., and Marillonnet, S., A One Pot, One Step, Precision Cloning Method with High Throughput Capability, *PLoS One* **3** (2008) e3647.

Grodberg, J. and Dunn, J.J., ompT encodes the Escherichia coli outer membrane protease that cleaves T7 RNA polymerase during purification, *J. Bacteriol.* **170** (1988) 1245–1253.

Gustafsson, C., Minshull, J., Govindarajan, S., Ness, J., Villalobos, A., and Welch, M., Engineering genes for predictable protein expression, *Protein Expr. Purif.* **83** (2012) 37–46.

Habibi, Y., Key advances in the chemical modification of nanocelluloses, *Chem. Soc. Rev.* **43** (2014) 1519–42.

Hu, H., Qian, J., Chu, J., Wang, Y., Zhuang, Y., and Zhang, S., DNA shuffling of methionine adenosyltransferase gene leads to improved S-adenosyl-l-methionine production in Pichia pastoris, *J. Biotechnol.* **141** (2009) 97–103.

Hu, S., Wang, M., Cai, G., and He, M., Genetic code-guided protein synthesis and folding in Escherichia coli, *J. Biol. Chem.* **288** (2013) 30855–30861.

Ikemura, T., Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer R, *J. Mol. Biol.* **158** (1982) 573–597.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., and Charpentier, E., A

Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity, *Science* **337** (2012) 816–821.

Jonsson, J., Norberg, T., Carlsson, L., Gustafsson, C., and Wold, S., Quantitative sequence-activity models (QSAM) - tools for sequence design, *Nucleic Acids Res.* **21** (1993) 733–9.

Kayano, K., Saruwatari, K., Kogure, T., and Shiraiwa, Y., Effect of Coccolith Polysaccharides Isolated from the Coccolithophorid, Emiliania huxleyi, on Calcite Crystal Formation in In Vitro $CaCO_3$ Crystallization, *Mar. Biotechnol.* **13** (2011) 83–92.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., and Drummond, A., Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data, *Bioinformatics* **28** (2012) 1647–1649.

Kraulis, J., Clore, G.M., Nilges, M., Jones, T.A., Pettersson, G., Knowles, J., and Gronenborn, A.M., Determination of the three-dimensional solution structure of the C-terminal domain of cellobiohydrolase I from Trichoderma reesei. A study using nuclear magnetic resonance and hybrid distance geometry-dynamical simulated annealing, *Biochemistry* **28** (1989) 7241–7257.

Kudla, G., Murray, A.W., Tollervey, D., and Plotkin, J.B., Coding-Sequence Determinants of Gene Expression in Escherichia coli, *Science* **324** (2009) 255–258.

Lander, E.S., The Heroes of CRISPR, *Cell* **164** (2016) 18–28.

Linder, M., *Structure-function relationships in fungal cellulose-binding domains*, Doctoral dissertation, VTT Publications 294, Espoo 1996, 29 p.

Linder, M., Mattinen, M.L., Kontteli, M., Lindeberg, G., Ståhlberg, J., Drakenberg, T., Reinikainen, T., Pettersson, G., and Annila, A., Identification of functionally

important amino acids in the cellulose-binding domain of Trichoderma reesei cellobiohydrolase I, *Protein Sci.* **4** (1995) 1056–1064.

Malho, J.-M., Arola, S., Laaksonen, P., Szilvay, G.R., Ikkala, O., and Linder, M.B., Modular Architecture of Protein Binding Units for Designing Properties of Cellulose Nanomaterials, *Angew. Chemie Int. Ed.* **54** (2015) 12025–12028.

Martinez, D., Berka, R.M., Henrissat, B., Saloheimo, M., Arvas, M., Baker, S.E., Chapman, J., Chertkov, O., Coutinho, P.M., Cullen, D., Danchin, E.G.J., Grigoriev, I. V, Harris, P., Jackson, M., Kubicek, C.P., Han, C.S., Ho, I., Larrondo, L.F., de Leon, A.L., Magnuson, J.K., Merino, S., Misra, M., Nelson, B., Putnam, N., Robbertse, B., Salamov, A.A., Schmoll, M., Terry, A., Thayer, N., Westerholm-Parvinen, A., Schoch, C.L., Yao, J., Barabote, R., Barbote, R., Nelson, M.A., Detter, C., Bruce, D., Kuske, C.R., Xie, G., Richardson, P., Rokhsar, D.S., Lucas, S.M., Rubin, E.M., Dunn-Coleman, N., Ward, M., and Brettin, T.S., Genome sequencing and analysis of the biomass-degrading fungus Trichoderma reesei (syn. Hypocrea jecorina), *Nat. Biotechnol.* **26** (2008) 553–560.

Matos, C.F.R.O., Robinson, C., Alanen, H.I., Prus, P., Uchida, Y., Ruddock, L.W., Freedman, R.B., and Keshavarz-Moore, E., Efficient export of prefolded, disulfide-bonded recombinant proteins to the periplasm by the Tat pathway in Escherichia coli CyDisCo strains, *Biotechnol. Prog.* **30** (2014) 281–290.

Moore, J.C. and Arnold, F.H., Directed evolution of a para-nitrobenzyl esterase for aqueous-organic solvents, *Nat Biotechnol* **14** (1996) 458–467.

New England Biolabs, Datasheet for BL21(DE3) Competent E. coli, https://www.neb.com/~/media/Catalog/All-Products/0B28021B9A36470BB46B318DAD19ED4F/Datacards%20or%20Manuals/C2527Datasheet-Lot22.pdf, 4 August 2016a.

New England Biolabs, Datasheet for T7 Express Competent E. coli, https://www.neb.com/~/media/Catalog/All-

Products/A9EA95B0D5B44ADB98CCB59560F65990/Datacards%20or%20Manuals/C2566Datasheet-Lot19.pdf, 4 August 2016b.

Novagen, *pET System Manual*, 11th ed., Billerica, MA, USA 2006, 63 p.

Omotajo, D., Tate, T., Cho, H., and Choudhary, M., Distribution and diversity of ribosome binding sites in prokaryotic genomes, *BMC Genomics* **16** (2015) 604.

Paasche, E., A review of the coccolithophorid Emiliania huxleyi (prymnesiophyceae), with particular reference to growth, coccolith formation, and calcification-photosynthesis interactions, *Phycologia* **40** (2001) 503–529.

Palonen, H., Tenkanen, M., and Linder, M., Dynamic interaction of Trichoderma reesei cellobiohydrolases Cel6A and Cel7A and cellulose at equilibrium and during hydrolysis, *Appl. Environ. Microbiol.* **65** (1999) 5229–5233.

Plotkin, J.B. and Kudla, G., Synonymous but not the same: the causes and consequences of codon bias, *Nat. Rev. Genet.* **12** (2011) 32–42.

Prinz, W.A., Åslund, F., Beckwith, J., and Holmgren, A., The Role of the Thioredoxin and Glutaredoxin Pathways in Reducing Protein Disulfide Bonds in the Escherichia coli Cytoplasm, *J. Biol. Chem.* **272** (1997) 15661–15667.

Read, B. a, Kegel, J., Klute, M.J., Kuo, A., Lefebvre, S.C., Maumus, F., Mayer, C., Miller, J., Monier, A., Salamov, A., Young, J., Aguilar, M., Claverie, J.-M., Frickenhaus, S., Gonzalez, K., Herman, E.K., Lin, Y.-C., Napier, J., Ogata, H., Sarno, A.F., Shmutz, J., Schroeder, D., de Vargas, C., Verret, F., von Dassow, P., Valentin, K., Van de Peer, Y., Wheeler, G., Dacks, J.B., Delwiche, C.F., Dyhrman, S.T., Glöckner, G., John, U., Richards, T., Worden, A.Z., Zhang, X., and Grigoriev, I. V, Pan genome of the phytoplankton Emiliania underpins its global distribution, *Nature* **499** (2013) 209–13.

Sharp, P.M. and Li, W.H., The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications, *Nucleic Acids Res.* **15**

(1987) 1281–1295.

Shetty, R.P., Endy, D., and Knight, T.F., Engineering BioBrick vectors from BioBrick parts, *J. Biol. Eng.* **2** (2008) 5.

Silva-Rocha, R., Martínez-García, E., Calles, B., Chavarría, M., Arce-Rodríguez, A., De Las Heras, A., Páez-Espino, A.D., Durante-Rodríguez, G., Kim, J., Nikel, P.I., Platero, R., and De Lorenzo, V., The Standard European Vector Architecture (SEVA): A coherent platform for the analysis and deployment of complex prokaryotic phenotypes, *Nucleic Acids Res.* **41** (2013) 666–675.

Stemmer, W.P., Rapid evolution of a protein in vitro by DNA shuffling, *Nature* **370** (1994a) 389–391.

Stemmer, W.P., DNA shuffling by random fragmentation and reassembly: in vitro recombination for molecular evolution, *Proc. Natl. Acad. Sci. U. S. A.* **91** (1994b) 10747–10751.

Studier, F.W., Use of bacteriophage T7 lysozyme to improve an inducible T7 expression system, *J. Mol. Biol.* **219** (1991) 37–44.

Studier, F.W., Protein production by auto-induction in high density shaking cultures, *Protein Expr. Purif.* **41** (2005) 207–234.

Studier, F.W. and Moffatt, B.A., Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes, *J. Mol. Biol.* **189** (1986) 113–130.

Tan, Y., Hoon, S., Guerette, P.A., Wei, W., Ghadban, A., Hao, C., Miserez, A., and Waite, J.H., Infiltration of chitin by protein coacervates defines the squid beak mechanical gradient, *Nat. Chem. Biol.* **11** (2015) 488–495.

Tsai, C.J., Sauna, Z.E., Kimchi-Sarfaty, C., Ambudkar, S. V., Gottesman, M.M., and Nussinov, R., Synonymous Mutations and Ribosome Stalling Can Lead to Altered Folding Pathways and Distinct Minima, *J. Mol. Biol.* **383** (2008) 281–291.

Welch, M., Govindarajan, S., Ness, J.E., Villalobos, A., Gurney, A., Minshull, J., and Gustafsson, C., Design Parameters to Control Synthetic Gene Expression in Escherichia coli, *PLoS One* **4** (2009a) e7002.

Welch, M., Villalobos, A., Gustafsson, C., and Minshull, J., You're one in a googol: optimizing genes for protein expression, *J. R. Soc. Interface* **6** (2009b) S467–S476.

Welch, M., Villalobos, A., Gustafsson, C., and Minshull, J., Designing genes for successful protein expression, *Methods Enzymol.* **498** (2011) 43–66.

Zetsche, B., Gootenberg, J.S., Abudayyeh, O.O., Slaymaker, I.M., Makarova, K.S., Essletzbichler, P., Volz, S.E., Joung, J., van der Oost, J., Regev, A., Koonin, E. V., and Zhang, F., Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System, *Cell* **163** (2015) 759–771.

Åslund, F. and Beckwith, J., The Thioredoxin Superfamily: Redundancy, Specificity, and Gray-Area Genomics, *J. Bacteriol.* **181** (1999) 1375–1379.

DNA sequences used in the experimental research

Primers

>MI02
AAGCATCGGTCTCGAGAGCCACCACCTTT
>MI03
AGCATCGGTCTCCATGAAATACCTATTGC
>pSVAR1
GATATAGGCGCCAGCAACC
>pSVAF2
CTTAATGCGCCGCTACAGG

Ordered gene fragments (GeneArt Strings DNA fragments, Life Technologies)

>Ehux1b2 (201 bp)
CTAGCTGGCTGAAGCTGACTGAGTTA**GGTCTC**CCTCTCCGAACCCCGGTCCTTGGGAACAGTGTGGTGGT
AAATCGTATGAAGGTCCGACCGCCTGTCCCCGTGAATACACCTGTCAGTATCGCCGTGAAACGTTTTCACA
GTGTATTCCTGGTGGTAAAGGCGAGC**GAGACC**CTGATAGCGGTATGACCTAGTAGCATGC

>Ehux1b2A (202 bp)
CTAGCTGGCTGAAGCTGACTGAGTTA**GGTCTC**CCTCTCCGACCCAGAGCCATTATGGTCAGTGTGGTGGTA
AATCGTATGAAGGTCCGACCGCCTGTCCCCGTGAATACACCTGTCAGTATCGCCGTGAAACGTTTTCACAG
TGTATTCCTGGTGGTAAAGGCGAGC**GAGACC**CTGATAGCGGTATGACCTAGTAGCATGCA

>Ehux1b2B (203 bp)
CTAGCTGGCTGAAGCTGACTGAGTTA**GGTCTC**CCTCTCCGAACCCCGGTCCTTGGGAACAGTGTGGTGGT
ATTGGTTATAGCGGTCCGACCGTTTGTCCCCGTGAATACACCTGTCAGTATCGCCGTGAAACGTTTTCACA
GTGTATTCCTGGTGGTAAAGGCGAGC**GAGACC**CTGATAGCGGTATGACCTAGTAGCATGCAT

>Ehux1b2C (204 bp)
CTAGCTGGCTGAAGCTGACTGAGTTA**GGTCTC**CCTCTCCGAACCCCGGTCCTTGGGAACAGTGTGGTGGT
AAATCGTATGAAGGTCCGACCGCCTGTGCAAGCGGCACCACCTGTCAGGTTCTGCGTGAAACGTTTTCACA
GTGTATTCCTGGTGGTAAAGGCGAGC**GAGACC**CTGATAGCGGTATGACCTAGTAGCATGCATC

>Ehux1b2D (205 bp)
CTAGCTGGCTGAAGCTGACTGAGTTA**GGTCTC**CCTCTCCGAACCCCGGTCCTTGGGAACAGTGTGGTGGT
AAATCGTATGAAGGTCCGACCGCCTGTCCCCGTGAATACACCTGTCAGTATCGCAATCCGTATTATTCACA
GTGTCTGCCTGGTGGTAAAGGCGAGC**GAGACC**CTGATAGCGGTATGACCTAGTAGCATGCATCC

>Cel7Awt (206 bp)
CTAGCTGGCTGAAGCTGACTGAGTTA**GGTCTC**CCTCTCCGACCCAGTCTCACTACGGCCAGTGCGGCGGTA
TTGGCTACAGCGGCCCCACGGTCTGCGCCAGCGGCACAACTTGCCAGGTCCTGAACCCTTACTACTCTCAG
TGCCTGCCTGGTGGTAAAGGCGAGC**GAGACC**CTGATAGCGGTATGACCTAGTAGCATGCATCCG

Existing fragments and the empty expression vector

>pelB/linker1 (152 bp)
GGTCTCCATGAAATACCTATTGCCTACGGCAGCCGCTGGATTGTTATTACTCGCGGCCCAGCCGGCCATGG
CATCTAC**GAGACC**AAAAGGGGCCCCTTTTAAAATTTT**GGTCTC**CGGCGGCGGTGGTAGCAAAGGTGGTGG
CTCTC**GAGACC**

>alkaline phosphatase (1371 bp)
GGTCTCATCTAGGACTCCGGAAATGCCGGTTCTGGAAAATCGTGCAGCACAGGGTGATATTACCGCACCG
GGTGGTGCACGTCGTCTGACCGGTGATCAGACCGCAGCACTGCGTGATAGCCTGAGCGATAAACCGGCA
AAAAACATTATTCTGCTGATTGGTGATGGCATGGGTGATAGCGAAATTACCGCAGCACGTAATTATGCCG
AAGGTGCCGGTGGTTTTTTTAAAGGTATTGATGCACTGCCGCTGACAGGTCAGTATACCCATTATGCACTG
AATAAAAAAACCGGCAAACCGGATTATGTTACCGATAGCGCAGCAAGCGCAACCGCATGGTCAACCGGTG
TTAAAACCTATAATGGTGCACTGGGTGTGGATATTCATGAAAAAGATCATCCGACCATTCTGGAAATGGCA
AAAGCAGCAGGTCTGGCAACCGGTAATGTTAGCACCGCAGAACTGCAGGATGCAACACCGGCAGCACTG
GTTGCACATGTTACCAGCCGTAAATGTTATGGTCCGAGCGCAACCAGCGAAAAATGTCCGGGTAATGCAC
TGGAAAAAGGTGGTAAAGGTAGCATTACCGAACAGCTGCTGAATGCACGTGCAGATGTTACCCTGGGTG
GTGGTGCAAAAACCTTTGCAGAAACCGCAACCGCAGGCGAATGGCAGGGTAAAACCCTGCGTGAACAGG
CACAGGCACGTGGTTATCAGCTGGTTAGTGATGCAGCAAGCCTGAATAGCGTTACCGAAGCAAATCAGCA
GAAACCGCTGCTGGGTCTGTTTGCAGATGGTAATATGCCGGTTCGTTGGCTGGGTCCGAAAGCAACCTAT
CATGGTAATATTGATAAACCGGCTGTTACCTGTACCCCGAATCCGCAGCGTAATGATAGCGTTCCGACCCT
GGCACAGATGACCGATAAAGCAATTGAACTGCTGAGCAAAAATGAAAAAGGCTTTTTTCTGCAGGTTGAA
GGTGCCAGCATTGATAAACAGGATCATGCAGCAAATCCGTGTGGTCAGATTGGTGAAACCGTTGATCTGG
ATGAAGCAGTTCAGCGTGCACTGGAATTTGCAAAAAAAGAAGGTAATACCCTGGTTATTGTGACCGCAGA
TCATGCACATGCAAGCCAGATTGTTGCACCGGATACCAAAGCACCGGGTCTGACCCAGGCACTGAATACC
AAAGATGGTGCAGTTATGGTTATGAGCTATGGTAATAGCGAAGAAGATAGCCAGGAACATACCGGTAGC
CAGCTGCGTATTGCAGCATATGGTCCGCATGCAGCCAATGTTGTTGGTCTGACCGATCAAACCGACCTGTT
TTATACCATGAAAGCAGCACTGGGTCTGAAAGGCGC**GAGACC**

>Empty expression plasmid pBR1a (5263 bp, pET28a+ derivative for Golden gate cloning)
TGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGAC
CGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGC
TTTCCCCGTCAAGCTCTAAATCGGGGGCTCCCTTTAGGGTTCCGATTTAGTGCTTTACGGCACCTCGACCCC
AAAAAACTTGATTAGGGTGATGGTTCACGTAGTGGGCCATCGCCCTGATAGACGGTTTTTCGCCCTTTGAC
GTTGGAGTCCACGTTCTTTAATAGTGGACTCTTGTTCCAAACTGGAACAACACTCAACCCTATCTCGGTCTA
TTCTTTTGATTTATAAGGGATTTTGCCGATTTCGGCCTATTGGTTAAAAAATGAGCTGATTTAACAAAAATT
TAACGCGAATTTTAACAAAATATTAACGCTTACAATTTAGGTGGCACTTTTCGGGGAAATGTGCGCGGAAC
CCCTATTTGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAATTAATTCTTAGAAAAACTCATC
GAGCATCAAATGAAACTGCAATTTATTCATATCAGGATTATCAATACCATATTTTTGAAAAAGCCGTTTCTG
TAATGAAGGAGAAAACTCACCGAGGCAGTTCCATAGGATGGCAAGATCCTGGTATCGGTCTGCGATTCCG
ACTCGTCCAACATCAATACAACCTATTAATTTCCCCTCGTCAAAAATAAGGTTATCAAGTGAGAAATCACCA
TGAGTGACGACTGAATCCGGTGAGAATGGCAAAAGTTTATGCATTTCTTTCCAGACTTGTTCAACAGGCCA
GCCATTACGCTCGTCATCAAAATCACTCGCATCAACCAAACCGTTATTCATTCGTGATTGCGCCTGAGCGA
GACGAAATACGCGATCGCTGTTAAAAGGACAATTACAAACAGGAATCGAATGCAACCGGCGCAGGAACA
CTGCCAGCGCATCAACAATATTTTCACCTGAATCAGGATATTCTTCTAATACCTGGAATGCTGTTTTCCCGG
GGATCGCAGTGGTGAGTAACCATGCATCATCAGGAGTACGGATAAAATGCTTGATGGTCGGAAGAGGCA
TAAATTCCGTCAGCCAGTTTAGTCTGACCATCTCATCTGTAACATCATTGGCAACGCTACCTTTGCCATGTTT
CAGAAACAACTCTGGCGCATCGGGCTTCCCATACAATCGATAGATTGTCGCACCTGATTGCCCGACATTAT
CGCGAGCCCATTTATACCCATATAAATCAGCATCCATGTTGGAATTTAATCGCGGCCTAGAGCAAGACGTT
TCCCGTTGAATATGGCTCATAACACCCCTTGTATTACTGTTTATGTAAGCAGACAGTTTTATTGTTCATGACC
AAAATCCCTTAACGTGAGTTTTCGTTCCACTGAGCGTCAGACCCCGTAGAAAAGATCAAAGGATCTTCTTG
AGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACCAGCGGTGGTTTGTTT
GCCGGATCAAGAGCTACCAACTCTTTTTCCGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACT
GTCCTTCTAGTGTAGCCGTAGTTAGGCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCTCTG

CTAATCCTGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAAGACGATA
GTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGGGTTCGTGCACACAGCCCAGCTTGGAGCGAAC
GACCTACACCGAACTGAGATACCTACAGCGTGAGCTATGAGAAAGCGCCACGCTTCCCGAAGGGAGAAA
GGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGA
AACGCCTGGTATCTTTATAGTCCTGTCGGGTTTCGCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCG
TCAGGGGGGCGGAGCCTATGGAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGC
CTTTTGCTCACATGTTCTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTATTACCGCCTTTGAGTGAGC
TGATACCGCTCGCCGCAGCCGAACGACCGAGCGCAGCGAGTCAGTGAGCGAGGAAGCGGAAGAGCGCC
TGATGCGGTATTTTCTCCTTACGCATCTGTGCGGTATTTCACACCGCATATATGGTGCACTCTCAGTACAAT
CTGCTCTGATGCCGCATAGTTAAGCCAGTATACACTCCGCTATCGCTACGTGACTGGGTCATGGCTGCGCC
CCGACACCCGCCAACACCCGCTGACGCGCCCTGACGGGCTTGTCTGCTCCCGGCATCCGCTTACAGACAAG
CTGTGACCGTCTCCGGGAGCTGCATGTGTCAGAGGTTTTCACCGTCATCACCGAAACGCGCGAGGCAGCT
GCGGTAAAGCTCATCAGCGTGGTCGTGAAGCGATTCACAGATGTCTGCCTGTTCATCCGCGTCCAGCTCGT
TGAGTTTCTCCAGAAGCGTTAATGTCTGGCTTCTGATAAAGCGGGCCATGTTAAGGGCGGTTTTTTCCTGT
TTGGTCACTGATGCCTCCGTGTAAGGGGGATTTCTGTTCATGGGGGTAATGATACCGATGAAACGAGAGA
GGATGCTCACGATACGGGTTACTGATGATGAACATGCCCGGTTACTGGAACGTTGTGAGGGTAAACAACT
GGCGGTATGGATGCGGCGGGACCAGAGAAAAATCACTCAGGGTCAATGCCAGCGCTTCGTTAATACAGA
TGTAGGTGTTCCACAGGGTAGCCAGCAGCATCCTGCGATGCAGATCCGGAACATAATGGTGCAGGGCGCT
GACTTCCGCGTTTCCAGACTTTACGAAACACGGAAACCGAAGACCATTCATGTTGTTGCTCAGGTCGCAGA
CGTTTTGCAGCAGCAGTCGCTTCACGTTCGCTCGCGTATCGGTGATTCATTCTGCTAACCAGTAAGGCAAC
CCCGCCAGCCTAGCCGGGTCCTCAACGACAGGAGCACGATCATGCGCACCCGTGGGGCCGCCATGCCGG
CGATAATGGCCTGCTTCTCGCCGAAACGTTTGGTGGCGGGACCAGTGACGAAGGCTTGAGCGAGGGCGT
GCAAGATTCCGAATACCGCAAGCGACAGGCCGATCATCGTCGCGCTCCAGCGAAAGCGGTCCTCGCCGAA
AATGACCCAGAGCGCTGCCGGCACCTGTCCTACGAGTTGCATGATAAAGAAGACAGTCATAAGTGCGGCG
ACGATAGTCATGCCCCGCGCCCACCGGAAGGAGCTGACTGGGTTGAAGGCTCTCAAGGGCATCGGTCGA
GATCCCGGTGCCTAATGAGTGAGCTAACTTACATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGG
AAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGCG
CCAGGGTGGTTTTTCTTTTCACCAGTGAGACGGGCAACAGCTGATTGCCCTTCACCGCCTGGCCCTGAGAG
AGTTGCAGCAAGCGGTCCACGCTGGTTTGCCCCAGCAGGCGAAAATCCTGTTTGATGGTGGTTAACGGCG
GGATATAACATGAGCTGTCTTCGGTATCGTCGTATCCCACTACCGAGATATCCGCACCAACGCGCAGCCCG
GACTCGGTAATGGCGCGCATTGCGCCCAGCGCCATCTGATCGTTGGCAACCAGCATCGCAGTGGGAACGA
TGCCCTCATTCAGCATTTGCATGGTTTGTTGAAAACCGGACATGGCACTCCAGTCGCCTTCCCGTTCCGCTA
TCGGCTGAATTTGATTGCGAGTGAGATATTTATGCCAGCCAGCCAGACGCAGACGCGCCGAGACAGAACT
TAATGGGCCCGCTAACAGCGCGATTTGCTGGTGACCCAATGCGACCAGATGCTCCACGCCCAGTCGCGTA
CCGTCTTCATGGGAGAAAATAATACTGTTGATGGGTGTCTGGTCAGAGACATCAAGAAATAACGCCGGAA
CATTAGTGCAGGCAGCTTCCACAGCAATGGCATCCTGGTCATCCAGCGGATAGTTAATGATCAGCCCACTG
ACGCGTTGCGCGAGAAGATTGTGCACCGCCGCTTTACAGGCTTCGACGCCGCTTCGTTCTACCATCGACAC
CACCACGCTGGCACCCAGTTGATCGGCGCGAGATTTAATCGCCGCGACAATTTGCGACGGCGCGTGCAGG
GCCAGACTGGAGGTGGCAACGCCAATCAGCAACGACTGTTTGCCCGCCAGTTGTTGTGCCACGCGGTTGG
GAATGTAATTCAGCTCCGCCATCGCCGCTTCCACTTTTTCCCGCGTTTTCGCAGAAACGTGGCTGGCCTGGT
TCACCACGCGGGAAACGGTCTGATAAGAGACACCGGCATACTCTGCGACATCGTATAACGTTACTGGTTTC
ACATTCACCACCCTGAATTGACTCTCTTCCGGGCGCTATCATGCCATACCGCGAAAGGTTTTGCGCCATTCG
ATGGTGTCCGGGATCTCGACGCTCTCCCTTATGCGACTCCTGCATTAGGAAGCAGCCCAGTAGTAGGTTGA
GGCCGTTGAGCACCGCCGCCGCAAGGAATGGTGCATGCAAGGAGATGGCGCCCAACAGTCCCCCGGCCA
CGGGGCCTGCCACCATACCCACGCCGAAACAAGCGCTCATGAGCCCGAAGTGGCGAGCCCGATCTTCCCC
ATCGGTGATGTCGGCGATATAGGCGCCAGCAACCGCACCTGTGGCGCCGGTGATGCCGGCCACGATGCG
TCCGGCGTAGAGGATCGAGATCTCGATCCCGCGAAATTAATACGACTCACTATAGGGGAATTGTGAGCGG
ATAACAATTCCCCTCTAGAAATAATTTTGTTTAACTTTAAGAAGGAGATATACCATGAG**GAGACC**GCGGAT
CCGAATTCG**GGTCTC**TCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGA
AGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTG
AGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGAT

Expression plasmids

>pelB-AP-Cel7Aopt._pBR1a (6804 bp)

TGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGAC
CGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGC
TTTCCCCGTCAAGCTCTAAATCGGGGGCTCCCTTTAGGGTTCCGATTTAGTGCTTTACGGCACCTCGACCCC
AAAAAACTTGATTAGGGTGATGGTTCACGTAGTGGGCCATCGCCCTGATAGACGGTTTTTCGCCCTTTGAC
GTTGGAGTCCACGTTCTTTAATAGTGGACTCTTGTTCCAAACTGGAACAACACTCAACCCTATCTCGGTCTA
TTCTTTTGATTTATAAGGGATTTTGCCGATTTCGGCCTATTGGTTAAAAAATGAGCTGATTTAACAAAAATT
TAACGCGAATTTTAACAAAATATTAACGCTTACAATTTAGGTGGCACTTTTCGGGGAAATGTGCGCGGAAC
CCCTATTTGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAATTAATTCTTAGAAAAACTCATC
GAGCATCAAATGAAACTGCAATTTATTCATATCAGGATTATCAATACCATATTTTTGAAAAAGCCGTTTCTG
TAATGAAGGAGAAAACTCACCGAGGCAGTTCCATAGGATGGCAAGATCCTGGTATCGGTCTGCGATTCCG
ACTCGTCCAACATCAATACAACCTATTAATTTCCCCTCGTCAAAAATAAGGTTATCAAGTGAGAAATCACCA
TGAGTGACGACTGAATCCGGTGAGAATGGCAAAAGTTTATGCATTTCTTTCCAGACTTGTTCAACAGGCCA
GCCATTACGCTCGTCATCAAAATCACTCGCATCAACCAAACCGTTATTCATTCGTGATTGCGCCTGAGCGA
GACGAAATACGCGATCGCTGTTAAAAGGACAATTACAAACAGGAATCGAATGCAACCGGCGCAGGAACA
CTGCCAGCGCATCAACAATATTTTCACCTGAATCAGGATATTCTTCTAATACCTGGAATGCTGTTTTCCCGG
GGATCGCAGTGGTGAGTAACCATGCATCATCAGGAGTACGGATAAAATGCTTGATGGTCGGAAGAGGCA
TAAATTCCGTCAGCCAGTTTAGTCTGACCATCTCATCTGTAACATCATTGGCAACGCTACCTTTGCCATGTTT
CAGAAACAACTCTGGCGCATCGGGCTTCCCATACAATCGATAGATTGTCGCACCTGATTGCCCGACATTAT
CGCGAGCCCATTTATACCCATATAAATCAGCATCCATGTTGGAATTTAATCGCGGCCTAGAGCAAGACGTT
TCCCGTTGAATATGGCTCATAACACCCCTTGTATTACTGTTTATGTAAGCAGACAGTTTTATTGTTCATGACC
AAAATCCCTTAACGTGAGTTTTCGTTCCACTGAGCGTCAGACCCCGTAGAAAAGATCAAAGGATCTTCTTG
AGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACCAGCGGTGGTTTGTTT
GCCGGATCAAGAGCTACCAACTCTTTTTCCGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACT
GTCCTTCTAGTGTAGCCGTAGTTAGGCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCTCTG
CTAATCCTGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAAGACGATA
GTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGGGTTCGTGCACACAGCCCAGCTTGGAGCGAAC
GACCTACACCGAACTGAGATACCTACAGCGTGAGCTATGAGAAAGCGCCACGCTTCCCGAAGGGAGAAA
GGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGA
AACGCCTGGTATCTTTATAGTCCTGTCGGGTTTCGCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCG
TCAGGGGGGCGGAGCCTATGGAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGC
CTTTTGCTCACATGTTCTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTATTACCGCCTTTGAGTGAGC
TGATACCGCTCGCCGCAGCCGAACGACCGAGCGCAGCGAGTCAGTGAGCGAGGAAGCGGAAGAGCGCC
TGATGCGGTATTTTCTCCTTACGCATCTGTGCGGTATTTCACACCGCATATATGGTGCACTCTCAGTACAAT
CTGCTCTGATGCCGCATAGTTAAGCCAGTATACACTCCGCTATCGCTACGTGACTGGGTCATGGCTGCGCC
CCGACACCCGCCAACACCCGCTGACGCGCCCTGACGGGCTTGTCTGCTCCCGGCATCCGCTTACAGACAAG
CTGTGACCGTCTCCGGGAGCTGCATGTGTCAGAGGTTTTCACCGTCATCACCGAAACGCGCGAGGCAGCT
GCGGTAAAGCTCATCAGCGTGGTCGTGAAGCGATTCACAGATGTCTGCCTGTTCATCCGCGTCCAGCTCGT
TGAGTTTCTCCAGAAGCGTTAATGTCTGGCTTCTGATAAAGCGGGCCATGTTAAGGGCGGTTTTTTCCTGT
TTGGTCACTGATGCCTCCGTGTAAGGGGGATTTCTGTTCATGGGGGTAATGATACCGATGAAACGAGAGA
GGATGCTCACGATACGGGTTACTGATGATGAACATGCCCGGTTACTGGAACGTTGTGAGGGTAAACAACT
GGCGGTATGGATGCGGCGGGACCAGAGAAAAATCACTCAGGGTCAATGCCAGCGCTTCGTTAATACAGA
TGTAGGTGTTCCACAGGGTAGCCAGCAGCATCCTGCGATGCAGATCCGGAACATAATGGTGCAGGGCGCT
GACTTCCGCGTTTCCAGACTTTACGAAACACGGAAACCGAAGACCATTCATGTTGTTGCTCAGGTCGCAGA
CGTTTTGCAGCAGCAGTCGCTTCACGTTCGCTCGCGTATCGGTGATTCATTCTGCTAACCAGTAAGGCAAC
CCCGCCAGCCTAGCCGGGTCCTCAACGACAGGAGCACGATCATGCGCACCCGTGGGGCCGCCATGCCGG
CGATAATGGCCTGCTTCTCGCCGAAACGTTTGGTGGCGGGACCAGTGACGAAGGCTTGAGCGAGGGCGT
GCAAGATTCCGAATACCGCAAGCGACAGGCCGATCATCGTCGCGCTCCAGCGAAAGCGGTCCTCGCCGAA
AATGACCCAGAGCGCTGCCGGCACCTGTCCTACGAGTTGCATGATAAAGAAGACAGTCATAAGTGCGGCG
ACGATAGTCATGCCCCGCGCCCACCGGAAGGAGCTGACTGGGTTGAAGGCTCTCAAGGGCATCGGTCGA
GATCCCGGTGCCTAATGAGTGAGCTAACTTACATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGG
AAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGCG

CCAGGGTGGTTTTTCTTTTCACCAGTGAGACGGGCAACAGCTGATTGCCCTTCACCGCCTGGCCCTGAGAG
AGTTGCAGCAAGCGGTCCACGCTGGTTTGCCCCAGCAGGCGAAAATCCTGTTTGATGGTGGTTAACGGCG
GGATATAACATGAGCTGTCTTCGGTATCGTCGTATCCCACTACCGAGATATCCGCACCAACGCGCAGCCCG
GACTCGGTAATGGCGCGCATTGCGCCCAGCGCCATCTGATCGTTGGCAACCAGCATCGCAGTGGGAACGA
TGCCCTCATTCAGCATTTGCATGGTTTGTTGAAAACCGGACATGGCACTCCAGTCGCCTTCCCGTTCCGCTA
TCGGCTGAATTTGATTGCGAGTGAGATATTTATGCCAGCCAGCCAGACGCAGACGCGCCGAGACAGAACT
TAATGGGCCCGCTAACAGCGCGATTTGCTGGTGACCCAATGCGACCAGATGCTCCACGCCCAGTCGCGTA
CCGTCTTCATGGGAGAAAATAATACTGTTGATGGGTGTCTGGTCAGAGACATCAAGAAATAACGCCGGAA
CATTAGTGCAGGCAGCTTCCACAGCAATGGCATCCTGGTCATCCAGCGGATAGTTAATGATCAGCCCACTG
ACGCGTTGCGCGAGAAGATTGTGCACCGCCGCTTTACAGGCTTCGACGCCGCTTCGTTCTACCATCGACAC
CACCACGCTGGCACCCAGTTGATCGGCGCGAGATTTAATCGCCGCGACAATTTGCGACGGCGCGTGCAGG
GCCAGACTGGAGGTGGCAACGCCAATCAGCAACGACTGTTTGCCCGCCAGTTGTTGTGCCACGCGGTTGG
GAATGTAATTCAGCTCCGCCATCGCCGCTTCCACTTTTTCCCGCGTTTTCGCAGAAACGTGGCTGGCCTGGT
TCACCACGCGGGAAACGGTCTGATAAGAGACACCGGCATACTCTGCGACATCGTATAACGTTACTGGTTTC
ACATTCACCACCCTGAATTGACTCTCTTCCGGGCGCTATCATGCCATACCGCGAAAGGTTTTGCGCCATTCG
ATGGTGTCCGGGATCTCGACGCTCTCCCTTATGCGACTCCTGCATTAGGAAGCAGCCCAGTAGTAGGTTGA
GGCCGTTGAGCACCGCCGCCGCAAGGAATGGTGCATGCAAGGAGATGGCGCCCAACAGTCCCCCGGCCA
CGGGGCCTGCCACCATACCCACGCCGAAACAAGCGCTCATGAGCCCGAAGTGGCGAGCCCGATCTTCCCC
ATCGGTGATGTCGGCGATATAGGCGCCAGCAACCGCACCTGTGGCGCCGGTGATGCCGGCCACGATGCG
TCCGGCGTAGAGGATCGAGATCTCGATCCCGCGAAATTAATACGACTCACTATAGGGGAATTGTGAGCGG
ATAACAATTCCCCTCTAGAAATAATTTTGTTTAACTTTAAGAAGGAGATATACCATGAAATACCTATTGCCT
ACGGCAGCCGCTGGATTGTTATTACTCGCGGCCCAGCCGGCCATGGCATCTAGGACTCCGGAAATGCCGG
TTCTGGAAAATCGTGCAGCACAGGGTGATATTACCGCACCGGGTGGTGCACGTCGTCTGACCGGTGATCA
GACCGCAGCACTGCGTGATAGCCTGAGCGATAAACCGGCAAAAAACATTATTCTGCTGATTGGTGATGGC
ATGGGTGATAGCGAAATTACCGCAGCACGTAATTATGCCGAAGGTGCCGGTGGTTTTTTTAAAGGTATTG
ATGCACTGCCGCTGACAGGTCAGTATACCCATTATGCACTGAATAAAAAAACCGGCAAACCGGATTATGTT
ACCGATAGCGCAGCAAGCGCAACCGCATGGTCAACCGGTGTTAAAACCTATAATGGTGCACTGGGTGTGG
ATATTCATGAAAAAGATCATCCGACCATTCTGGAAATGGCAAAAGCAGCAGGTCTGGCAACCGGTAATGT
TAGCACCGCAGAACTGCAGGATGCAACACCGGCAGCACTGGTTGCACATGTTACCAGCCGTAAATGTTAT
GGTCCGAGCGCAACCAGCGAAAAATGTCCGGGTAATGCACTGGAAAAAGGTGGTAAAGGTAGCATTACC
GAACAGCTGCTGAATGCACGTGCAGATGTTACCCTGGGTGGTGGTGCAAAAACCTTTGCAGAAACCGCAA
CCGCAGGCGAATGGCAGGGTAAAACCCTGCGTGAACAGGCACAGGCACGTGGTTATCAGCTGGTTAGTG
ATGCAGCAAGCCTGAATAGCGTTACCGAAGCAAATCAGCAGAAACCGCTGCTGGGTCTGTTTGCAGATGG
TAATATGCCGGTTCGTTGGCTGGGTCCGAAAGCAACCTATCATGGTAATATTGATAAACCGGCTGTTACCT
GTACCCCGAATCCGCAGCGTAATGATAGCGTTCCGACCCTGGCACAGATGACCGATAAAGCAATTGAACT
GCTGAGCAAAAATGAAAAAGGCTTTTTTCTGCAGGTTGAAGGTGCCAGCATTGATAAACAGGATCATGCA
GCAAATCCGTGTGGTCAGATTGGTGAAACCGTTGATCTGGATGAAGCAGTTCAGCGTGCACTGGAATTTG
CAAAAAAAGAAGGTAATACCCTGGTTATTGTGACCGCAGATCATGCACATGCAAGCCAGATTGTTGCACC
GGATACCAAAGCACCGGGTCTGACCCAGGCACTGAATACCAAAGATGGTGCAGTTATGGTTATGAGCTAT
GGTAATAGCGAAGAAGATAGCCAGGAACATACCGGTAGCCAGCTGCGTATTGCAGCATATGGTCCGCAT
GCAGCCAATGTTGTTGGTCTGACCGATCAAACCGACCTGTTTTATACCATGAAAGCAGCACTGGGTCTGAA
AGGCGGCGGTGGTAGCAAAGGTGGTGGCTCTCCGACCCAGAGCCATTATGGTCAGTGTGGTGGTATTGG
TTATAGCGGTCCGACCGTTTGTGCAAGCGGCACCACCTGTCAGGTTCTGAATCCGTATTATTCACAGTGTCT
GCCTGGTGGTAAAGGCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGA
AGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTG
AGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGAT

> pelB-AP-Cel7Awt-pBR1a (6804 bp)

TGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGAC
CGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGC
TTTCCCCGTCAAGCTCTAAATCGGGGGCTCCCTTTAGGGTTCCGATTTAGTGCTTTACGGCACCTCGACCCC
AAAAAACTTGATTAGGGTGATGGTTCACGTAGTGGGCCATCGCCCTGATAGACGGTTTTTCGCCCTTTGAC
GTTGGAGTCCACGTTCTTTAATAGTGGACTCTTGTTCCAAACTGGAACAACACTCAACCCTATCTCGGTCTA
TTCTTTTGATTTATAAGGGATTTTGCCGATTTCGGCCTATTGGTTAAAAAATGAGCTGATTTAACAAAAATT

TAACGCGAATTTTAACAAAATATTAACGCTTACAATTTAGGTGGCACTTTTCGGGGAAATGTGCGCGGAAC
CCCTATTTGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAATTAATTCTTAGAAAAACTCATC
GAGCATCAAATGAAACTGCAATTTATTCATATCAGGATTATCAATACCATATTTTTGAAAAAGCCGTTTCTG
TAATGAAGGAGAAAACTCACCGAGGCAGTTCCATAGGATGGCAAGATCCTGGTATCGGTCTGCGATTCCG
ACTCGTCCAACATCAATACAACCTATTAATTTCCCCTCGTCAAAAATAAGGTTATCAAGTGAGAAATCACCA
TGAGTGACGACTGAATCCGGTGAGAATGGCAAAAGTTTATGCATTTCTTTCCAGACTTGTTCAACAGGCCA
GCCATTACGCTCGTCATCAAAATCACTCGCATCAACCAAACCGTTATTCATTCGTGATTGCGCCTGAGCGA
GACGAAATACGCGATCGCTGTTAAAAGGACAATTACAAACAGGAATCGAATGCAACCGGCGCAGGAACA
CTGCCAGCGCATCAACAATATTTTCACCTGAATCAGGATATTCTTCTAATACCTGGAATGCTGTTTTCCCGG
GGATCGCAGTGGTGAGTAACCATGCATCATCAGGAGTACGGATAAAATGCTTGATGGTCGGAAGAGGCA
TAAATTCCGTCAGCCAGTTTAGTCTGACCATCTCATCTGTAACATCATTGGCAACGCTACCTTTGCCATGTTT
CAGAAACAACTCTGGCGCATCGGGCTTCCCATACAATCGATAGATTGTCGCACCTGATTGCCCGACATTAT
CGCGAGCCCATTTATACCCATATAAATCAGCATCCATGTTGGAATTTAATCGCGGCCTAGAGCAAGACGTT
TCCCGTTGAATATGGCTCATAACACCCCTTGTATTACTGTTTATGTAAGCAGACAGTTTTATTGTTCATGACC
AAAATCCCTTAACGTGAGTTTTCGTTCCACTGAGCGTCAGACCCCGTAGAAAAGATCAAAGGATCTTCTTG
AGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACCAGCGGTGGTTTGTTT
GCCGGATCAAGAGCTACCAACTCTTTTTCCGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACT
GTCCTTCTAGTGTAGCCGTAGTTAGGCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCTCTG
CTAATCCTGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAAGACGATA
GTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGGGTTCGTGCACACAGCCCAGCTTGGAGCGAAC
GACCTACACCGAACTGAGATACCTACAGCGTGAGCTATGAGAAAGCGCCACGCTTCCCGAAGGGAGAAA
GGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGA
AACGCCTGGTATCTTTATAGTCCTGTCGGGTTTCGCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCG
TCAGGGGGGCGGAGCCTATGGAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGC
CTTTTGCTCACATGTTCTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTATTACCGCCTTTGAGTGAGC
TGATACCGCTCGCCGCAGCCGAACGACCGAGCGCAGCGAGTCAGTGAGCGAGGAAGCGGAAGAGCGCC
TGATGCGGTATTTTCTCCTTACGCATCTGTGCGGTATTTCACACCGCATATATGGTGCACTCTCAGTACAAT
CTGCTCTGATGCCGCATAGTTAAGCCAGTATACACTCCGCTATCGCTACGTGACTGGGTCATGGCTGCGCC
CCGACACCCGCCAACACCCGCTGACGCGCCCTGACGGGCTTGTCTGCTCCCGGCATCCGCTTACAGACAAG
CTGTGACCGTCTCCGGGAGCTGCATGTGTCAGAGGTTTTCACCGTCATCACCGAAACGCGCGAGGCAGCT
GCGGTAAAGCTCATCAGCGTGGTCGTGAAGCGATTCACAGATGTCTGCCTGTTCATCCGCGTCCAGCTCGT
TGAGTTTCTCCAGAAGCGTTAATGTCTGGCTTCTGATAAAGCGGGCCATGTTAAGGGCGGTTTTTTCCTGT
TTGGTCACTGATGCCTCCGTGTAAGGGGGATTTCTGTTCATGGGGGTAATGATACCGATGAAACGAGAGA
GGATGCTCACGATACGGGTTACTGATGATGAACATGCCCGGTTACTGGAACGTTGTGAGGGTAAACAACT
GGCGGTATGGATGCGGCGGGACCAGAGAAAAATCACTCAGGGTCAATGCCAGCGCTTCGTTAATACAGA
TGTAGGTGTTCCACAGGGTAGCCAGCAGCATCCTGCGATGCAGATCCGGAACATAATGGTGCAGGGCGCT
GACTTCCGCGTTTCCAGACTTTACGAAACACGGAAACCGAAGACCATTCATGTTGTTGCTCAGGTCGCAGA
CGTTTTGCAGCAGCAGTCGCTTCACGTTCGCTCGCGTATCGGTGATTCATTCTGCTAACCAGTAAGGCAAC
CCCGCCAGCCTAGCCGGGTCCTCAACGACAGGAGCACGATCATGCGCACCCGTGGGGCCGCCATGCCGG
CGATAATGGCCTGCTTCTCGCCGAAACGTTTGGTGGCGGGACCAGTGACGAAGGCTTGAGCGAGGGCGT
GCAAGATTCCGAATACCGCAAGCGACAGGCCGATCATCGTCGCGCTCCAGCGAAAGCGGTCCTCGCCGAA
AATGACCCAGAGCGCTGCCGGCACCTGTCCTACGAGTTGCATGATAAAGAAGACAGTCATAAGTGCGGCG
ACGATAGTCATGCCCCGCGCCCACCGGAAGGAGCTGACTGGGTTGAAGGCTCTCAAGGGCATCGGTCGA
GATCCCGGTGCCTAATGAGTGAGCTAACTTACATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGG
AAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGCG
CCAGGGTGGTTTTTCTTTTCACCAGTGAGACGGGCAACAGCTGATTGCCCTTCACCGCCTGGCCCTGAGAG
AGTTGCAGCAAGCGGTCCACGCTGGTTTGCCCCAGCAGGCGAAAATCCTGTTTGATGGTGGTTAACGGCG
GGATATAACATGAGCTGTCTTCGGTATCGTCGTATCCCACTACCGAGATATCCGCACCAACGCGCAGCCCG
GACTCGGTAATGGCGCGCATTGCGCCCAGCGCCATCTGATCGTTGGCAACCAGCATCGCAGTGGGAACGA
TGCCCTCATTCAGCATTTGCATGGTTTGTTGAAAACCGGACATGGCACTCCAGTCGCCTTCCCGTTCCGCTA
TCGGCTGAATTTGATTGCGAGTGAGATATTTATGCCAGCCAGCCAGACGCAGACGCGCCGAGACAGAACT
TAATGGGCCCGCTAACAGCGCGATTTGCTGGTGACCCAATGCGACCAGATGCTCCACGCCCAGTCGCGTA
CCGTCTTCATGGGAGAAAATAATACTGTTGATGGGTGTCTGGTCAGAGACATCAAGAAATAACGCCGGAA
CATTAGTGCAGGCAGCTTCCACAGCAATGGCATCCTGGTCATCCAGCGGATAGTTAATGATCAGCCCACTG
ACGCGTTGCGCGAGAAGATTGTGCACCGCCGCTTTACAGGCTTCGACGCCGCTTCGTTCTACCATCGACAC

CACCACGCTGGCACCCAGTTGATCGGCGCGAGATTTAATCGCCGCGACAATTTGCGACGGCGCGTGCAGG
GCCAGACTGGAGGTGGCAACGCCAATCAGCAACGACTGTTTGCCCGCCAGTTGTTGTGCCACGCGGTTGG
GAATGTAATTCAGCTCCGCCATCGCCGCTTCCACTTTTTCCCGCGTTTTCGCAGAAACGTGGCTGGCCTGGT
TCACCACGCGGGAAACGGTCTGATAAGAGACACCGGCATACTCTGCGACATCGTATAACGTTACTGGTTTC
ACATTCACCACCCTGAATTGACTCTCTTCCGGGCGCTATCATGCCATACCGCGAAAGGTTTTGCGCCATTCG
ATGGTGTCCGGGATCTCGACGCTCTCCCTTATGCGACTCCTGCATTAGGAAGCAGCCCAGTAGTAGGTTGA
GGCCGTTGAGCACCGCCGCCGCAAGGAATGGTGCATGCAAGGAGATGGCGCCCAACAGTCCCCCGGCCA
CGGGGCCTGCCACCATACCCACGCCGAAACAAGCGCTCATGAGCCCGAAGTGGCGAGCCCGATCTTCCCC
ATCGGTGATGTCGGCGATATAGGCGCCAGCAACCGCACCTGTGGCGCCGGTGATGCCGGCCACGATGCG
TCCGGCGTAGAGGATCGAGATCTCGATCCCGCGAAATTAATACGACTCACTATAGGGGAATTGTGAGCGG
ATAACAATTCCCCTCTAGAAATAATTTTGTTTAACTTTAAGAAGGAGATATACCATGAAATACCTATTGCCT
ACGGCAGCCGCTGGATTGTTATTACTCGCGGCCCAGCCGGCCATGGCATCTAGGACTCCGGAAATGCCGG
TTCTGGAAAATCGTGCAGCACAGGGTGATATTACCGCACCGGGTGGTGCACGTCGTCTGACCGGTGATCA
GACCGCAGCACTGCGTGATAGCCTGAGCGATAAACCGGCAAAAAACATTATTCTGCTGATTGGTGATGGC
ATGGGTGATAGCGAAATTACCGCAGCACGTAATTATGCCGAAGGTGCCGGTGGTTTTTTTAAAGGTATTG
ATGCACTGCCGCTGACAGGTCAGTATACCCATTATGCACTGAATAAAAAAACCGGCAAACCGGATTATGTT
ACCGATAGCGCAGCAAGCGCAACCGCATGGTCAACCGGTGTTAAAACCTATAATGGTGCACTGGGTGTGG
ATATTCATGAAAAAGATCATCCGACCATTCTGGAAATGGCAAAAGCAGCAGGTCTGGCAACCGGTAATGT
TAGCACCGCAGAACTGCAGGATGCAACACCGGCAGCACTGGTTGCACATGTTACCAGCCGTAAATGTTAT
GGTCCGAGCGCAACCAGCGAAAAATGTCCGGGTAATGCACTGGAAAAAGGTGGTAAAGGTAGCATTACC
GAACAGCTGCTGAATGCACGTGCAGATGTTACCCTGGGTGGTGGTGCAAAAACCTTTGCAGAAACCGCAA
CCGCAGGCGAATGGCAGGGTAAAACCCTGCGTGAACAGGCACAGGCACGTGGTTATCAGCTGGTTAGTG
ATGCAGCAAGCCTGAATAGCGTTACCGAAGCAAATCAGCAGAAACCGCTGCTGGGTCTGTTTGCAGATGG
TAATATGCCGGTTCGTTGGCTGGGTCCGAAAGCAACCTATCATGGTAATATTGATAAACCGGCTGTTACCT
GTACCCCGAATCCGCAGCGTAATGATAGCGTTCCGACCCTGGCACAGATGACCGATAAAGCAATTGAACT
GCTGAGCAAAAATGAAAAAGGCTTTTTTCTGCAGGTTGAAGGTGCCAGCATTGATAAACAGGATCATGCA
GCAAATCCGTGTGGTCAGATTGGTGAAACCGTTGATCTGGATGAAGCAGTTCAGCGTGCACTGGAATTTG
CAAAAAAAGAAGGTAATACCCTGGTTATTGTGACCGCAGATCATGCACATGCAAGCCAGATTGTTGCACC
GGATACCAAAGCACCGGGTCTGACCCAGGCACTGAATACCAAAGATGGTGCAGTTATGGTTATGAGCTAT
GGTAATAGCGAAGAAGATAGCCAGGAACATACCGGTAGCCAGCTGCGTATTGCAGCATATGGTCCGCAT
GCAGCCAATGTTGTTGGTCTGACCGATCAAACCGACCTGTTTTATACCATGAAAGCAGCACTGGGTCTGAA
AGGCGGCGGTGGTAGCAAAGGTGGTGGCTCTCCGACCCAGTCTCACTACGGCCAGTGCGGCGGTATTGG
CTACAGCGGCCCCACGGTCTGCGCCAGCGGCACAACTTGCCAGGTCCTGAACCCTTACTACTCTCAGTGCC
TGCCTGGTGGTAAAGGCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGA
AGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTG
AGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGAT

>pelB-AP-Ehux1b2 -pBR1a (6804 bp)

TGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGAC
CGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGC
TTTCCCCGTCAAGCTCTAAATCGGGGGCTCCCTTTAGGGTTCCGATTTAGTGCTTTACGGCACCTCGACCCC
AAAAAACTTGATTAGGGTGATGGTTCACGTAGTGGGCCATCGCCCTGATAGACGGTTTTTCGCCCTTTGAC
GTTGGAGTCCACGTTCTTTAATAGTGGACTCTTGTTCCAAACTGGAACAACACTCAACCCTATCTCGGTCTA
TTCTTTTGATTTATAAGGGATTTTGCCGATTTCGGCCTATTGGTTAAAAAATGAGCTGATTTAACAAAAATT
TAACGCGAATTTTAACAAAATATTAACGCTTACAATTTAGGTGGCACTTTTCGGGGAAATGTGCGCGGAAC
CCCTATTTGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAATTAATTCTTAGAAAAACTCATC
GAGCATCAAATGAAACTGCAATTTATTCATATCAGGATTATCAATACCATATTTTTGAAAAAGCCGTTTCTG
TAATGAAGGAGAAAACTCACCGAGGCAGTTCCATAGGATGGCAAGATCCTGGTATCGGTCTGCGATTCCG
ACTCGTCCAACATCAATACAACCTATTAATTTCCCCTCGTCAAAAATAAGGTTATCAAGTGAGAAATCACCA
TGAGTGACGACTGAATCCGGTGAGAATGGCAAAAGTTTATGCATTTCTTTCCAGACTTGTTCAACAGGCCA
GCCATTACGCTCGTCATCAAAATCACTCGCATCAACCAAACCGTTATTCATTCGTGATTGCGCCTGAGCGA
GACGAAATACGCGATCGCTGTTAAAAGGACAATTACAAACAGGAATCGAATGCAACCGGCGCAGGAACA
CTGCCAGCGCATCAACAATATTTTCACCTGAATCAGGATATTCTTCTAATACCTGGAATGCTGTTTTCCCGG
GGATCGCAGTGGTGAGTAACCATGCATCATCAGGAGTACGGATAAAATGCTTGATGGTCGGAAGAGGCA

TAAATTCCGTCAGCCAGTTTAGTCTGACCATCTCATCTGTAACATCATTGGCAACGCTACCTTTGCCATGTTT
CAGAAACAACTCTGGCGCATCGGGCTTCCCATACAATCGATAGATTGTCGCACCTGATTGCCCGACATTAT
CGCGAGCCCATTTATACCCATATAAATCAGCATCCATGTTGGAATTTAATCGCGGCCTAGAGCAAGACGTT
TCCCGTTGAATATGGCTCATAACACCCCTTGTATTACTGTTTATGTAAGCAGACAGTTTTATTGTTCATGACC
AAAATCCCTTAACGTGAGTTTTCGTTCCACTGAGCGTCAGACCCCGTAGAAAAGATCAAAGGATCTTCTTG
AGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACCAGCGGTGGTTTGTTT
GCCGGATCAAGAGCTACCAACTCTTTTTCCGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACT
GTCCTTCTAGTGTAGCCGTAGTTAGGCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCTCTG
CTAATCCTGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAAGACGATA
GTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGGGTTCGTGCACACAGCCCAGCTTGGAGCGAAC
GACCTACACCGAACTGAGATACCTACAGCGTGAGCTATGAGAAAGCGCCACGCTTCCCGAAGGGAGAAA
GGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGA
AACGCCTGGTATCTTTATAGTCCTGTCGGGTTTCGCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCG
TCAGGGGGGCGGAGCCTATGGAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGC
CTTTTGCTCACATGTTCTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTATTACCGCCTTTGAGTGAGC
TGATACCGCTCGCCGCAGCCGAACGACCGAGCGCAGCGAGTCAGTGAGCGAGGAAGCGGAAGAGCGCC
TGATGCGGTATTTTCTCCTTACGCATCTGTGCGGTATTTCACACCGCATATATGGTGCACTCTCAGTACAAT
CTGCTCTGATGCCGCATAGTTAAGCCAGTATACACTCCGCTATCGCTACGTGACTGGGTCATGGCTGCGCC
CCGACACCCGCCAACACCCGCTGACGCGCCCTGACGGGCTTGTCTGCTCCCGGCATCCGCTTACAGACAAG
CTGTGACCGTCTCCGGGAGCTGCATGTGTCAGAGGTTTTCACCGTCATCACCGAAACGCGCGAGGCAGCT
GCGGTAAAGCTCATCAGCGTGGTCGTGAAGCGATTCACAGATGTCTGCCTGTTCATCCGCGTCCAGCTCGT
TGAGTTTCTCCAGAAGCGTTAATGTCTGGCTTCTGATAAAGCGGGCCATGTTAAGGGCGGTTTTTTCCTGT
TTGGTCACTGATGCCTCCGTGTAAGGGGGATTTCTGTTCATGGGGGTAATGATACCGATGAAACGAGAGA
GGATGCTCACGATACGGGTTACTGATGATGAACATGCCCGGTTACTGGAACGTTGTGAGGGTAAACAACT
GGCGGTATGGATGCGGCGGGACCAGAGAAAAATCACTCAGGGTCAATGCCAGCGCTTCGTTAATACAGA
TGTAGGTGTTCCACAGGGTAGCCAGCAGCATCCTGCGATGCAGATCCGGAACATAATGGTGCAGGGCGCT
GACTTCCGCGTTTCCAGACTTTACGAAACACGGAAACCGAAGACCATTCATGTTGTTGCTCAGGTCGCAGA
CGTTTTGCAGCAGCAGTCGCTTCACGTTCGCTCGCGTATCGGTGATTCATTCTGCTAACCAGTAAGGCAAC
CCCGCCAGCCTAGCCGGGTCCTCAACGACAGGAGCACGATCATGCGCACCCGTGGGGCCGCCATGCCGG
CGATAATGGCCTGCTTCTCGCCGAAACGTTTGGTGGCGGGACCAGTGACGAAGGCTTGAGCGAGGGCGT
GCAAGATTCCGAATACCGCAAGCGACAGGCCGATCATCGTCGCGCTCCAGCGAAAGCGGTCCTCGCCGAA
AATGACCCAGAGCGCTGCCGGCACCTGTCCTACGAGTTGCATGATAAAGAAGACAGTCATAAGTGCGGCG
ACGATAGTCATGCCCCGCGCCCACCGGAAGGAGCTGACTGGGTTGAAGGCTCTCAAGGGCATCGGTCGA
GATCCCGGTGCCTAATGAGTGAGCTAACTTACATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGG
AAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGCG
CCAGGGTGGTTTTTCTTTTCACCAGTGAGACGGGCAACAGCTGATTGCCCTTCACCGCCTGGCCCTGAGAG
AGTTGCAGCAAGCGGTCCACGCTGGTTTGCCCCAGCAGGCGAAAATCCTGTTTGATGGTGGTTAACGGCG
GGATATAACATGAGCTGTCTTCGGTATCGTCGTATCCCACTACCGAGATATCCGCACCAACGCGCAGCCCG
GACTCGGTAATGGCGCGCATTGCGCCCAGCGCCATCTGATCGTTGGCAACCAGCATCGCAGTGGGAACGA
TGCCCTCATTCAGCATTTGCATGGTTTGTTGAAAACCGGACATGGCACTCCAGTCGCCTTCCCGTTCCGCTA
TCGGCTGAATTTGATTGCGAGTGAGATATTTATGCCAGCCAGCCAGACGCAGACGCGCCGAGACAGAACT
TAATGGGCCCGCTAACAGCGCGATTTGCTGGTGACCCAATGCGACCAGATGCTCCACGCCCAGTCGCGTA
CCGTCTTCATGGGAGAAAATAATACTGTTGATGGGTGTCTGGTCAGAGACATCAAGAAATAACGCCGGAA
CATTAGTGCAGGCAGCTTCCACAGCAATGGCATCCTGGTCATCCAGCGGATAGTTAATGATCAGCCCACTG
ACGCGTTGCGCGAGAAGATTGTGCACCGCCGCTTTACAGGCTTCGACGCCGCTTCGTTCTACCATCGACAC
CACCACGCTGGCACCCAGTTGATCGGCGCGAGATTTAATCGCCGCGACAATTTGCGACGGCGCGTGCAGG
GCCAGACTGGAGGTGGCAACGCCAATCAGCAACGACTGTTTGCCCGCCAGTTGTTGTGCCACGCGGTTGG
GAATGTAATTCAGCTCCGCCATCGCCGCTTCCACTTTTTCCCGCGTTTTCGCAGAAACGTGGCTGGCCTGGT
TCACCACGCGGGAAACGGTCTGATAAGAGACACCGGCATACTCTGCGACATCGTATAACGTTACTGGTTTC
ACATTCACCACCCTGAATTGACTCTCTTCCGGGCGCTATCATGCCATACCGCGAAAGGTTTTGCGCCATTCG
ATGGTGTCCGGGATCTCGACGCTCTCCCTTATGCGACTCCTGCATTAGGAAGCAGCCCAGTAGTAGGTTGA
GGCCGTTGAGCACCGCCGCCGCAAGGAATGGTGCATGCAAGGAGATGGCGCCCAACAGTCCCCCGGCCA
CGGGGCCTGCCACCATACCCACGCCGAAACAAGCGCTCATGAGCCCGAAGTGGCGAGCCCGATCTTCCCC
ATCGGTGATGTCGGCGATATAGGCGCCAGCAACCGCACCTGTGGCGCCGGTGATGCCGGCCACGATGCG
TCCGGCGTAGAGGATCGAGATCTCGATCCCGCGAAATTAATACGACTCACTATAGGGGAATTGTGAGCGG

ATAACAATTCCCCTCTAGAAATAATTTTGTTTAACTTTAAGAAGGAGATATACCATGAAATACCTATTGCCT
ACGGCAGCCGCTGGATTGTTATTACTCGCGGCCCAGCCGGCCATGGCATCTAGGACTCCGGAAATGCCGG
TTCTGGAAAATCGTGCAGCACAGGGTGATATTACCGCACCGGGTGGTGCACGTCGTCTGACCGGTGATCA
GACCGCAGCACTGCGTGATAGCCTGAGCGATAAACCGGCAAAAAACATTATTCTGCTGATTGGTGATGGC
ATGGGTGATAGCGAAATTACCGCAGCACGTAATTATGCCGAAGGTGCCGGTGGTTTTTTTAAAGGTATTG
ATGCACTGCCGCTGACAGGTCAGTATACCCATTATGCACTGAATAAAAAAACCGGCAAACCGGATTATGTT
ACCGATAGCGCAGCAAGCGCAACCGCATGGTCAACCGGTGTTAAAACCTATAATGGTGCACTGGGTGTGG
ATATTCATGAAAAAGATCATCCGACCATTCTGGAAATGGCAAAAGCAGCAGGTCTGGCAACCGGTAATGT
TAGCACCGCAGAACTGCAGGATGCAACACCGGCAGCACTGGTTGCACATGTTACCAGCCGTAAATGTTAT
GGTCCGAGCGCAACCAGCGAAAAATGTCCGGGTAATGCACTGGAAAAAGGTGGTAAAGGTAGCATTACC
GAACAGCTGCTGAATGCACGTGCAGATGTTACCCTGGGTGGTGGTGCAAAAACCTTTGCAGAAACCGCAA
CCGCAGGCGAATGGCAGGGTAAAACCCTGCGTGAACAGGCACAGGCACGTGGTTATCAGCTGGTTAGTG
ATGCAGCAAGCCTGAATAGCGTTACCGAAGCAAATCAGCAGAAACCGCTGCTGGGTCTGTTTGCAGATGG
TAATATGCCGGTTCGTTGGCTGGGTCCGAAAGCAACCTATCATGGTAATATTGATAAACCGGCTGTTACCT
GTACCCCGAATCCGCAGCGTAATGATAGCGTTCCGACCCTGGCACAGATGACCGATAAAGCAATTGAACT
GCTGAGCAAAAATGAAAAAGGCTTTTTTCTGCAGGTTGAAGGTGCCAGCATTGATAAACAGGATCATGCA
GCAAATCCGTGTGGTCAGATTGGTGAAACCGTTGATCTGGATGAAGCAGTTCAGCGTGCACTGGAATTTG
CAAAAAAAGAAGGTAATACCCTGGTTATTGTGACCGCAGATCATGCACATGCAAGCCAGATTGTTGCACC
GGATACCAAAGCACCGGGTCTGACCCAGGCACTGAATACCAAAGATGGTGCAGTTATGGTTATGAGCTAT
GGTAATAGCGAAGAAGATAGCCAGGAACATACCGGTAGCCAGCTGCGTATTGCAGCATATGGTCCGCAT
GCAGCCAATGTTGTTGGTCTGACCGATCAAACCGACCTGTTTTATACCATGAAAGCAGCACTGGGTCTGAA
AGGCGGCGGTGGTAGCAAAGGTGGTGGCTCTCCGAACCCCGGTCCTTGGGAACAGTGTGGTGGTAAATC
GTATGAAGGTCCGACCGCCTGTCCCCGTGAATACACCTGTCAGTATCGCCGTGAAACGTTTTCACAGTGTA
TTCCTGGTGGTAAAGGCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGA
AGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTG
AGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGAT

>pelB-AP-Ehux1b2A -pBR1a (6804 bp)

TGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGAC
CGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGC
TTTCCCCGTCAAGCTCTAAATCGGGGGCTCCCTTTAGGGTTCCGATTTAGTGCTTTACGGCACCTCGACCCC
AAAAAACTTGATTAGGGTGATGGTTCACGTAGTGGGCCATCGCCCTGATAGACGGTTTTTCGCCCTTTGAC
GTTGGAGTCCACGTTCTTTAATAGTGGACTCTTGTTCCAAACTGGAACAACACTCAACCCTATCTCGGTCTA
TTCTTTTGATTTATAAGGGATTTTGCCGATTTCGGCCTATTGGTTAAAAAATGAGCTGATTTAACAAAAATT
TAACGCGAATTTTAACAAAATATTAACGCTTACAATTTAGGTGGCACTTTTCGGGGAAATGTGCGCGGAAC
CCCTATTTGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAATTAATTCTTAGAAAAACTCATC
GAGCATCAAATGAAACTGCAATTTATTCATATCAGGATTATCAATACCATATTTTTGAAAAAGCCGTTTCTG
TAATGAAGGAGAAAACTCACCGAGGCAGTTCCATAGGATGGCAAGATCCTGGTATCGGTCTGCGATTCCG
ACTCGTCCAACATCAATACAACCTATTAATTTCCCCTCGTCAAAAATAAGGTTATCAAGTGAGAAATCACCA
TGAGTGACGACTGAATCCGGTGAGAATGGCAAAAGTTTATGCATTTCTTTCCAGACTTGTTCAACAGGCCA
GCCATTACGCTCGTCATCAAAATCACTCGCATCAACCAAACCGTTATTCATTCGTGATTGCGCCTGAGCGA
GACGAAATACGCGATCGCTGTTAAAAGGACAATTACAAACAGGAATCGAATGCAACCGGCGCAGGAACA
CTGCCAGCGCATCAACAATATTTTCACCTGAATCAGGATATTCTTCTAATACCTGGAATGCTGTTTTCCCGG
GGATCGCAGTGGTGAGTAACCATGCATCATCAGGAGTACGGATAAAATGCTTGATGGTCGGAAGAGGCA
TAAATTCCGTCAGCCAGTTTAGTCTGACCATCTCATCTGTAACATCATTGGCAACGCTACCTTTGCCATGTTT
CAGAAACAACTCTGGCGCATCGGGCTTCCCATACAATCGATAGATTGTCGCACCTGATTGCCCGACATTAT
CGCGAGCCCATTTATACCCATATAAATCAGCATCCATGTTGGAATTTAATCGCGGCCTAGAGCAAGACGTT
TCCCGTTGAATATGGCTCATAACACCCCTTGTATTACTGTTTATGTAAGCAGACAGTTTTATTGTTCATGACC
AAAATCCCTTAACGTGAGTTTTCGTTCCACTGAGCGTCAGACCCCGTAGAAAAGATCAAAGGATCTTCTTG
AGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACCAGCGGTGGTTTGTTT
GCCGGATCAAGAGCTACCAACTCTTTTTCCGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACT
GTCCTTCTAGTGTAGCCGTAGTTAGGCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCTCTG
CTAATCCTGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAAGACGATA
GTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGGGTTCGTGCACACAGCCCAGCTTGGAGCGAAC

GACCTACACCGAACTGAGATACCTACAGCGTGAGCTATGAGAAAGCGCCACGCTTCCCGAAGGGAGAAA
GGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGA
AACGCCTGGTATCTTTATAGTCCTGTCGGGTTTCGCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCG
TCAGGGGGGCGGAGCCTATGGAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGC
CTTTTGCTCACATGTTCTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTATTACCGCCTTTGAGTGAGC
TGATACCGCTCGCCGCAGCCGAACGACCGAGCGCAGCGAGTCAGTGAGCGAGGAAGCGGAAGAGCGCC
TGATGCGGTATTTTCTCCTTACGCATCTGTGCGGTATTTCACACCGCATATATGGTGCACTCTCAGTACAAT
CTGCTCTGATGCCGCATAGTTAAGCCAGTATACACTCCGCTATCGCTACGTGACTGGGTCATGGCTGCGCC
CCGACACCCGCCAACACCCGCTGACGCGCCCTGACGGGCTTGTCTGCTCCCGGCATCCGCTTACAGACAAG
CTGTGACCGTCTCCGGGAGCTGCATGTGTCAGAGGTTTTCACCGTCATCACCGAAACGCGCGAGGCAGCT
GCGGTAAAGCTCATCAGCGTGGTCGTGAAGCGATTCACAGATGTCTGCCTGTTCATCCGCGTCCAGCTCGT
TGAGTTTCTCCAGAAGCGTTAATGTCTGGCTTCTGATAAAGCGGGCCATGTTAAGGGCGGTTTTTTCCTGT
TTGGTCACTGATGCCTCCGTGTAAGGGGGATTTCTGTTCATGGGGGTAATGATACCGATGAAACGAGAGA
GGATGCTCACGATACGGGTTACTGATGATGAACATGCCCGGTTACTGGAACGTTGTGAGGGTAAACAACT
GGCGGTATGGATGCGGCGGGACCAGAGAAAAATCACTCAGGGTCAATGCCAGCGCTTCGTTAATACAGA
TGTAGGTGTTCCACAGGGTAGCCAGCAGCATCCTGCGATGCAGATCCGGAACATAATGGTGCAGGGCGCT
GACTTCCGCGTTTCCAGACTTTACGAAACACGGAAACCGAAGACCATTCATGTTGTTGCTCAGGTCGCAGA
CGTTTTGCAGCAGCAGTCGCTTCACGTTCGCTCGCGTATCGGTGATTCATTCTGCTAACCAGTAAGGCAAC
CCCGCCAGCCTAGCCGGGTCCTCAACGACAGGAGCACGATCATGCGCACCCGTGGGGCCGCCATGCCGG
CGATAATGGCCTGCTTCTCGCCGAAACGTTTGGTGGCGGGACCAGTGACGAAGGCTTGAGCGAGGGCGT
GCAAGATTCCGAATACCGCAAGCGACAGGCCGATCATCGTCGCGCTCCAGCGAAAGCGGTCCTCGCCGAA
AATGACCCAGAGCGCTGCCGGCACCTGTCCTACGAGTTGCATGATAAAGAAGACAGTCATAAGTGCGGCG
ACGATAGTCATGCCCCGCGCCCACCGGAAGGAGCTGACTGGGTTGAAGGCTCTCAAGGGCATCGGTCGA
GATCCCGGTGCCTAATGAGTGAGCTAACTTACATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGG
AAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGCG
CCAGGGTGGTTTTTCTTTTCACCAGTGAGACGGGCAACAGCTGATTGCCCTTCACCGCCTGGCCCTGAGAG
AGTTGCAGCAAGCGGTCCACGCTGGTTTGCCCCAGCAGGCGAAAATCCTGTTTGATGGTGGTTAACGGCG
GGATATAACATGAGCTGTCTTCGGTATCGTCGTATCCCACTACCGAGATATCCGCACCAACGCGCAGCCCG
GACTCGGTAATGGCGCGCATTGCGCCCAGCGCCATCTGATCGTTGGCAACCAGCATCGCAGTGGGAACGA
TGCCCTCATTCAGCATTTGCATGGTTTGTTGAAAACCGGACATGGCACTCCAGTCGCCTTCCCGTTCCGCTA
TCGGCTGAATTTGATTGCGAGTGAGATATTTATGCCAGCCAGCCAGACGCAGACGCGCCGAGACAGAACT
TAATGGGCCCGCTAACAGCGCGATTTGCTGGTGACCCAATGCGACCAGATGCTCCACGCCCAGTCGCGTA
CCGTCTTCATGGGAGAAAATAATACTGTTGATGGGTGTCTGGTCAGAGACATCAAGAAATAACGCCGGAA
CATTAGTGCAGGCAGCTTCCACAGCAATGGCATCCTGGTCATCCAGCGGATAGTTAATGATCAGCCCACTG
ACGCGTTGCGCGAGAAGATTGTGCACCGCCGCTTTACAGGCTTCGACGCCGCTTCGTTCTACCATCGACAC
CACCACGCTGGCACCCAGTTGATCGGCGCGAGATTTAATCGCCGCGACAATTTGCGACGGCGCGTGCAGG
GCCAGACTGGAGGTGGCAACGCCAATCAGCAACGACTGTTTGCCCGCCAGTTGTTGTGCCACGCGGTTGG
GAATGTAATTCAGCTCCGCCATCGCCGCTTCCACTTTTTCCCGCGTTTTCGCAGAAACGTGGCTGGCCTGGT
TCACCACGCGGGAAACGGTCTGATAAGAGACACCGGCATACTCTGCGACATCGTATAACGTTACTGGTTTC
ACATTCACCACCCTGAATTGACTCTCTTCCGGGCGCTATCATGCCATACCGCGAAAGGTTTTGCGCCATTCG
ATGGTGTCCGGGATCTCGACGCTCTCCCTTATGCGACTCCTGCATTAGGAAGCAGCCCAGTAGTAGGTTGA
GGCCGTTGAGCACCGCCGCCGCAAGGAATGGTGCATGCAAGGAGATGGCGCCCAACAGTCCCCCGGCCA
CGGGGCCTGCCACCATACCCACGCCGAAACAAGCGCTCATGAGCCCGAAGTGGCGAGCCCGATCTTCCCC
ATCGGTGATGTCGGCGATATAGGCGCCAGCAACCGCACCTGTGGCGCCGGTGATGCCGGCCACGATGCG
TCCGGCGTAGAGGATCGAGATCTCGATCCCGCGAAATTAATACGACTCACTATAGGGGAATTGTGAGCGG
ATAACAATTCCCCTCTAGAAATAATTTTGTTTAACTTTAAGAAGGAGATATACCATGAAATACCTATTGCCT
ACGGCAGCCGCTGGATTGTTATTACTCGCGGCCCAGCCGGCCATGGCATCTAGGACTCCGGAAATGCCGG
TTCTGGAAAATCGTGCAGCACAGGGTGATATTACCGCACCGGGTGGTGCACGTCGTCTGACCGGTGATCA
GACCGCAGCACTGCGTGATAGCCTGAGCGATAAACCGGCAAAAAACATTATTCTGCTGATTGGTGATGGC
ATGGGTGATAGCGAAATTACCGCAGCACGTAATTATGCCGAAGGTGCCGGTGGTTTTTTTAAAGGTATTG
ATGCACTGCCGCTGACAGGTCAGTATACCCATTATGCACTGAATAAAAAAACCGGCAAACCGGATTATGTT
ACCGATAGCGCAGCAAGCGCAACCGCATGGTCAACCGGTGTTAAAACCTATAATGGTGCACTGGGTGTGG
ATATTCATGAAAAAGATCATCCGACCATTCTGGAAATGGCAAAAGCAGCAGGTCTGGCAACCGGTAATGT
TAGCACCGCAGAACTGCAGGATGCAACACCGGCAGCACTGGTTGCACATGTTACCAGCCGTAAATGTTAT
GGTCCGAGCGCAACCAGCGAAAAATGTCCGGGTAATGCACTGGAAAAAGGTGGTAAAGGTAGCATTACC

GAACAGCTGCTGAATGCACGTGCAGATGTTACCCTGGGTGGTGGTGCAAAAACCTTTGCAGAAACCGCAA
CCGCAGGCGAATGGCAGGGTAAAACCCTGCGTGAACAGGCACAGGCACGTGGTTATCAGCTGGTTAGTG
ATGCAGCAAGCCTGAATAGCGTTACCGAAGCAAATCAGCAGAAACCGCTGCTGGGTCTGTTTGCAGATGG
TAATATGCCGGTTCGTTGGCTGGGTCCGAAAGCAACCTATCATGGTAATATTGATAAACCGGCTGTTACCT
GTACCCCGAATCCGCAGCGTAATGATAGCGTTCCGACCCTGGCACAGATGACCGATAAAGCAATTGAACT
GCTGAGCAAAAATGAAAAAGGCTTTTTTCTGCAGGTTGAAGGTGCCAGCATTGATAAACAGGATCATGCA
GCAAATCCGTGTGGTCAGATTGGTGAAACCGTTGATCTGGATGAAGCAGTTCAGCGTGCACTGGAATTTG
CAAAAAAAGAAGGTAATACCCTGGTTATTGTGACCGCAGATCATGCACATGCAAGCCAGATTGTTGCACC
GGATACCAAAGCACCGGGTCTGACCCAGGCACTGAATACCAAAGATGGTGCAGTTATGGTTATGAGCTAT
GGTAATAGCGAAGAAGATAGCCAGGAACATACCGGTAGCCAGCTGCGTATTGCAGCATATGGTCCGCAT
GCAGCCAATGTTGTTGGTCTGACCGATCAAACCGACCTGTTTTATACCATGAAAGCAGCACTGGGTCTGAA
AGGCGGCGGTGGTAGCAAAGGTGGTGGCTCTCCGACCCAGAGCCATTATGGTCAGTGTGGTGGTAAATC
GTATGAAGGTCCGACCGCCTGTCCCCGTGAATACACCTGTCAGTATCGCCGTGAAACGTTTTCACAGTGTA
TTCCTGGTGGTAAAGGCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGA
AGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTG
AGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGAT

>pelB-AP-Ehux1b2B-pBR1a (6804 bp)

TGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGAC
CGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGC
TTTCCCCGTCAAGCTCTAAATCGGGGGCTCCCTTTAGGGTTCCGATTTAGTGCTTTACGGCACCTCGACCCC
AAAAAACTTGATTAGGGTGATGGTTCACGTAGTGGGCCATCGCCCTGATAGACGGTTTTTCGCCCTTTGAC
GTTGGAGTCCACGTTCTTTAATAGTGGACTCTTGTTCCAAACTGGAACAACACTCAACCCTATCTCGGTCTA
TTCTTTTGATTTATAAGGGATTTTGCCGATTTCGGCCTATTGGTTAAAAAATGAGCTGATTTAACAAAAATT
TAACGCGAATTTTAACAAAATATTAACGCTTACAATTTAGGTGGCACTTTTCGGGGAAATGTGCGCGGAAC
CCCTATTTGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAATTAATTCTTAGAAAAACTCATC
GAGCATCAAATGAAACTGCAATTTATTCATATCAGGATTATCAATACCATATTTTTGAAAAAGCCGTTTCTG
TAATGAAGGAGAAAACTCACCGAGGCAGTTCCATAGGATGGCAAGATCCTGGTATCGGTCTGCGATTCCG
ACTCGTCCAACATCAATACAACCTATTAATTTCCCCTCGTCAAAAATAAGGTTATCAAGTGAGAAATCACCA
TGAGTGACGACTGAATCCGGTGAGAATGGCAAAAGTTTATGCATTTCTTTCCAGACTTGTTCAACAGGCCA
GCCATTACGCTCGTCATCAAAATCACTCGCATCAACCAAACCGTTATTCATTCGTGATTGCGCCTGAGCGA
GACGAAATACGCGATCGCTGTTAAAAGGACAATTACAAACAGGAATCGAATGCAACCGGCGCAGGAACA
CTGCCAGCGCATCAACAATATTTTCACCTGAATCAGGATATTCTTCTAATACCTGGAATGCTGTTTTCCCGG
GGATCGCAGTGGTGAGTAACCATGCATCATCAGGAGTACGGATAAAATGCTTGATGGTCGGAAGAGGCA
TAAATTCCGTCAGCCAGTTTAGTCTGACCATCTCATCTGTAACATCATTGGCAACGCTACCTTTGCCATGTTT
CAGAAACAACTCTGGCGCATCGGGCTTCCCATACAATCGATAGATTGTCGCACCTGATTGCCCGACATTAT
CGCGAGCCCATTTATACCCATATAAATCAGCATCCATGTTGGAATTTAATCGCGGCCTAGAGCAAGACGTT
TCCCGTTGAATATGGCTCATAACACCCCTTGTATTACTGTTTATGTAAGCAGACAGTTTTATTGTTCATGACC
AAAATCCCTTAACGTGAGTTTTCGTTCCACTGAGCGTCAGACCCCGTAGAAAAGATCAAAGGATCTTCTTG
AGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACCAGCGGTGGTTTGTTT
GCCGGATCAAGAGCTACCAACTCTTTTTCCGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACT
GTCCTTCTAGTGTAGCCGTAGTTAGGCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCTCTG
CTAATCCTGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAAGACGATA
GTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGGGTTCGTGCACACAGCCCAGCTTGGAGCGAAC
GACCTACACCGAACTGAGATACCTACAGCGTGAGCTATGAGAAAGCGCCACGCTTCCCGAAGGGAGAAA
GGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGA
AACGCCTGGTATCTTTATAGTCCTGTCGGGTTTCGCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCG
TCAGGGGGGCGGAGCCTATGGAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGC
CTTTTGCTCACATGTTCTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTATTACCGCCTTTGAGTGAGC
TGATACCGCTCGCCGCAGCCGAACGACCGAGCGCAGCGAGTCAGTGAGCGAGGAAGCGGAAGAGCGCC
TGATGCGGTATTTTCTCCTTACGCATCTGTGCGGTATTTCACACCGCATATATGGTGCACTCTCAGTACAAT
CTGCTCTGATGCCGCATAGTTAAGCCAGTATACACTCCGCTATCGCTACGTGACTGGGTCATGGCTGCGCC
CCGACACCCGCCAACACCCGCTGACGCGCCCTGACGGGCTTGTCTGCTCCCGGCATCCGCTTACAGACAAG
CTGTGACCGTCTCCGGGAGCTGCATGTGTCAGAGGTTTTCACCGTCATCACCGAAACGCGCGAGGCAGCT

GCGGTAAAGCTCATCAGCGTGGTCGTGAAGCGATTCACAGATGTCTGCCTGTTCATCCGCGTCCAGCTCGT
TGAGTTTCTCCAGAAGCGTTAATGTCTGGCTTCTGATAAAGCGGGCCATGTTAAGGGCGGTTTTTTCCTGT
TTGGTCACTGATGCCTCCGTGTAAGGGGGATTTCTGTTCATGGGGGTAATGATACCGATGAAACGAGAGA
GGATGCTCACGATACGGGTTACTGATGATGAACATGCCCGGTTACTGGAACGTTGTGAGGGTAAACAACT
GGCGGTATGGATGCGGCGGGACCAGAGAAAAATCACTCAGGGTCAATGCCAGCGCTTCGTTAATACAGA
TGTAGGTGTTCCACAGGGTAGCCAGCAGCATCCTGCGATGCAGATCCGGAACATAATGGTGCAGGGCGCT
GACTTCCGCGTTTCCAGACTTTACGAAACACGGAAACCGAAGACCATTCATGTTGTTGCTCAGGTCGCAGA
CGTTTTGCAGCAGCAGTCGCTTCACGTTCGCTCGCGTATCGGTGATTCATTCTGCTAACCAGTAAGGCAAC
CCCGCCAGCCTAGCCGGGTCCTCAACGACAGGAGCACGATCATGCGCACCCGTGGGGCCGCCATGCCGG
CGATAATGGCCTGCTTCTCGCCGAAACGTTTGGTGGCGGGACCAGTGACGAAGGCTTGAGCGAGGGCGT
GCAAGATTCCGAATACCGCAAGCGACAGGCCGATCATCGTCGCGCTCCAGCGAAAGCGGTCCTCGCCGAA
AATGACCCAGAGCGCTGCCGGCACCTGTCCTACGAGTTGCATGATAAAGAAGACAGTCATAAGTGCGGCG
ACGATAGTCATGCCCCGCGCCCACCGGAAGGAGCTGACTGGGTTGAAGGCTCTCAAGGGCATCGGTCGA
GATCCCGGTGCCTAATGAGTGAGCTAACTTACATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGG
AAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGCG
CCAGGGTGGTTTTTCTTTTCACCAGTGAGACGGGCAACAGCTGATTGCCCTTCACCGCCTGGCCCTGAGAG
AGTTGCAGCAAGCGGTCCACGCTGGTTTGCCCCAGCAGGCGAAAATCCTGTTTGATGGTGGTTAACGGCG
GGATATAACATGAGCTGTCTTCGGTATCGTCGTATCCCACTACCGAGATATCCGCACCAACGCGCAGCCCG
GACTCGGTAATGGCGCGCATTGCGCCCAGCGCCATCTGATCGTTGGCAACCAGCATCGCAGTGGGAACGA
TGCCCTCATTCAGCATTTGCATGGTTTGTTGAAAACCGGACATGGCACTCCAGTCGCCTTCCCGTTCCGCTA
TCGGCTGAATTTGATTGCGAGTGAGATATTTATGCCAGCCAGCCAGACGCAGACGCGCCGAGACAGAACT
TAATGGGCCCGCTAACAGCGCGATTTGCTGGTGACCCAATGCGACCAGATGCTCCACGCCCAGTCGCGTA
CCGTCTTCATGGGAGAAAATAATACTGTTGATGGGTGTCTGGTCAGAGACATCAAGAAATAACGCCGGAA
CATTAGTGCAGGCAGCTTCCACAGCAATGGCATCCTGGTCATCCAGCGGATAGTTAATGATCAGCCCACTG
ACGCGTTGCGCGAGAAGATTGTGCACCGCCGCTTTACAGGCTTCGACGCCGCTTCGTTCTACCATCGACAC
CACCACGCTGGCACCCAGTTGATCGGCGCGAGATTTAATCGCCGCGACAATTTGCGACGGCGCGTGCAGG
GCCAGACTGGAGGTGGCAACGCCAATCAGCAACGACTGTTTGCCCGCCAGTTGTTGTGCCACGCGGTTGG
GAATGTAATTCAGCTCCGCCATCGCCGCTTCCACTTTTTCCCGCGTTTTCGCAGAAACGTGGCTGGCCTGGT
TCACCACGCGGGAAACGGTCTGATAAGAGACACCGGCATACTCTGCGACATCGTATAACGTTACTGGTTTC
ACATTCACCACCCTGAATTGACTCTCTTCCGGGCGCTATCATGCCATACCGCGAAAGGTTTTGCGCCATTCG
ATGGTGTCCGGGATCTCGACGCTCTCCCTTATGCGACTCCTGCATTAGGAAGCAGCCCAGTAGTAGGTTGA
GGCCGTTGAGCACCGCCGCCGCAAGGAATGGTGCATGCAAGGAGATGGCGCCCAACAGTCCCCCGGCCA
CGGGGCCTGCCACCATACCCACGCCGAAACAAGCGCTCATGAGCCCGAAGTGGCGAGCCCGATCTTCCCC
ATCGGTGATGTCGGCGATATAGGCGCCAGCAACCGCACCTGTGGCGCCGGTGATGCCGGCCACGATGCG
TCCGGCGTAGAGGATCGAGATCTCGATCCCGCGAAATTAATACGACTCACTATAGGGGAATTGTGAGCGG
ATAACAATTCCCCTCTAGAAATAATTTTGTTTAACTTTAAGAAGGAGATATACCATGAAATACCTATTGCCT
ACGGCAGCCGCTGGATTGTTATTACTCGCGGCCCAGCCGGCCATGGCATCTAGGACTCCGGAAATGCCGG
TTCTGGAAAATCGTGCAGCACAGGGTGATATTACCGCACCGGGTGGTGCACGTCGTCTGACCGGTGATCA
GACCGCAGCACTGCGTGATAGCCTGAGCGATAAACCGGCAAAAAACATTATTCTGCTGATTGGTGATGGC
ATGGGTGATAGCGAAATTACCGCAGCACGTAATTATGCCGAAGGTGCCGGTGGTTTTTTTAAAGGTATTG
ATGCACTGCCGCTGACAGGTCAGTATACCCATTATGCACTGAATAAAAAAACCGGCAAACCGGATTATGTT
ACCGATAGCGCAGCAAGCGCAACCGCATGGTCAACCGGTGTTAAAACCTATAATGGTGCACTGGGTGTGG
ATATTCATGAAAAAGATCATCCGACCATTCTGGAAATGGCAAAAGCAGCAGGTCTGGCAACCGGTAATGT
TAGCACCGCAGAACTGCAGGATGCAACACCGGCAGCACTGGTTGCACATGTTACCAGCCGTAAATGTTAT
GGTCCGAGCGCAACCAGCGAAAAATGTCCGGGTAATGCACTGGAAAAAGGTGGTAAAGGTAGCATTACC
GAACAGCTGCTGAATGCACGTGCAGATGTTACCCTGGGTGGTGGTGCAAAAACCTTTGCAGAAACCGCAA
CCGCAGGCGAATGGCAGGGTAAAACCCTGCGTGAACAGGCACAGGCACGTGGTTATCAGCTGGTTAGTG
ATGCAGCAAGCCTGAATAGCGTTACCGAAGCAAATCAGCAGAAACCGCTGCTGGGTCTGTTTGCAGATGG
TAATATGCCGGTTCGTTGGCTGGGTCCGAAAGCAACCTATCATGGTAATATTGATAAACCGGCTGTTACCT
GTACCCCGAATCCGCAGCGTAATGATAGCGTTCCGACCCTGGCACAGATGACCGATAAAGCAATTGAACT
GCTGAGCAAAAATGAAAAAGGCTTTTTTCTGCAGGTTGAAGGTGCCAGCATTGATAAACAGGATCATGCA
GCAAATCCGTGTGGTCAGATTGGTGAAACCGTTGATCTGGATGAAGCAGTTCAGCGTGCACTGGAATTTG
CAAAAAAAGAAGGTAATACCCTGGTTATTGTGACCGCAGATCATGCACATGCAAGCCAGATTGTTGCACC
GGATACCAAAGCACCGGGTCTGACCCAGGCACTGAATACCAAAGATGGTGCAGTTATGGTTATGAGCTAT
GGTAATAGCGAAGAAGATAGCCAGGAACATACCGGTAGCCAGCTGCGTATTGCAGCATATGGTCCGCAT

GCAGCCAATGTTGTTGGTCTGACCGATCAAACCGACCTGTTTTATACCATGAAAGCAGCACTGGGTCTGAA
AGGCGGCGGTGGTAGCAAAGGTGGTGGCTCTCCGAACCCCGGTCCTTGGGAACAGTGTGGTGGTATTGG
TTATAGCGGTCCGACCGTTTGTCCCCGTGAATACACCTGTCAGTATCGCCGTGAAACGTTTTCACAGTGTAT
TCCTGGTGGTAAAGGCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAA
GCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGA
GGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGAT

>pelB-AP-Ehux1b2C-pBR1a (6804 bp)
TGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGAC
CGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGC
TTTCCCCGTCAAGCTCTAAATCGGGGGCTCCCTTTAGGGTTCCGATTTAGTGCTTTACGGCACCTCGACCCC
AAAAAACTTGATTAGGGTGATGGTTCACGTAGTGGGCCATCGCCCTGATAGACGGTTTTTCGCCCTTTGAC
GTTGGAGTCCACGTTCTTTAATAGTGGACTCTTGTTCCAAACTGGAACAACACTCAACCCTATCTCGGTCTA
TTCTTTTGATTTATAAGGGATTTTGCCGATTTCGGCCTATTGGTTAAAAAATGAGCTGATTTAACAAAAATT
TAACGCGAATTTTAACAAAATATTAACGCTTACAATTTAGGTGGCACTTTTCGGGGAAATGTGCGCGGAAC
CCCTATTTGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAATTAATTCTTAGAAAAACTCATC
GAGCATCAAATGAAACTGCAATTTATTCATATCAGGATTATCAATACCATATTTTTGAAAAAGCCGTTTCTG
TAATGAAGGAGAAAACTCACCGAGGCAGTTCCATAGGATGGCAAGATCCTGGTATCGGTCTGCGATTCCG
ACTCGTCCAACATCAATACAACCTATTAATTTCCCCTCGTCAAAAATAAGGTTATCAAGTGAGAAATCACCA
TGAGTGACGACTGAATCCGGTGAGAATGGCAAAAGTTTATGCATTTCTTTCCAGACTTGTTCAACAGGCCA
GCCATTACGCTCGTCATCAAAATCACTCGCATCAACCAAACCGTTATTCATTCGTGATTGCGCCTGAGCGA
GACGAAATACGCGATCGCTGTTAAAAGGACAATTACAAACAGGAATCGAATGCAACCGGCGCAGGAACA
CTGCCAGCGCATCAACAATATTTTCACCTGAATCAGGATATTCTTCTAATACCTGGAATGCTGTTTTCCCGG
GGATCGCAGTGGTGAGTAACCATGCATCATCAGGAGTACGGATAAAATGCTTGATGGTCGGAAGAGGCA
TAAATTCCGTCAGCCAGTTTAGTCTGACCATCTCATCTGTAACATCATTGGCAACGCTACCTTTGCCATGTTT
CAGAAACAACTCTGGCGCATCGGGCTTCCCATACAATCGATAGATTGTCGCACCTGATTGCCCGACATTAT
CGCGAGCCCATTTATACCCATATAAATCAGCATCCATGTTGGAATTTAATCGCGGCCTAGAGCAAGACGTT
TCCCGTTGAATATGGCTCATAACACCCCTTGTATTACTGTTTATGTAAGCAGACAGTTTTATTGTTCATGACC
AAAATCCCTTAACGTGAGTTTTCGTTCCACTGAGCGTCAGACCCCGTAGAAAAGATCAAAGGATCTTCTTG
AGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACCAGCGGTGGTTTGTTT
GCCGGATCAAGAGCTACCAACTCTTTTTCCGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACT
GTCCTTCTAGTGTAGCCGTAGTTAGGCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCTCTG
CTAATCCTGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAAGACGATA
GTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGGGTTCGTGCACACAGCCCAGCTTGGAGCGAAC
GACCTACACCGAACTGAGATACCTACAGCGTGAGCTATGAGAAAGCGCCACGCTTCCCGAAGGGAGAAA
GGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGA
AACGCCTGGTATCTTTATAGTCCTGTCGGGTTTCGCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCG
TCAGGGGGGCGGAGCCTATGGAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGC
CTTTTGCTCACATGTTCTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTATTACCGCCTTTGAGTGAGC
TGATACCGCTCGCCGCAGCCGAACGACCGAGCGCAGCGAGTCAGTGAGCGAGGAAGCGGAAGAGCGCC
TGATGCGGTATTTTCTCCTTACGCATCTGTGCGGTATTTCACACCGCATATATGGTGCACTCTCAGTACAAT
CTGCTCTGATGCCGCATAGTTAAGCCAGTATACACTCCGCTATCGCTACGTGACTGGGTCATGGCTGCGCC
CCGACACCCGCCAACACCCGCTGACGCGCCCTGACGGGCTTGTCTGCTCCCGGCATCCGCTTACAGACAAG
CTGTGACCGTCTCCGGGAGCTGCATGTGTCAGAGGTTTTCACCGTCATCACCGAAACGCGCGAGGCAGCT
GCGGTAAAGCTCATCAGCGTGGTCGTGAAGCGATTCACAGATGTCTGCCTGTTCATCCGCGTCCAGCTCGT
TGAGTTTCTCCAGAAGCGTTAATGTCTGGCTTCTGATAAAGCGGGCCATGTTAAGGGCGGTTTTTTCCTGT
TTGGTCACTGATGCCTCCGTGTAAGGGGGATTTCTGTTCATGGGGGTAATGATACCGATGAAACGAGAGA
GGATGCTCACGATACGGGTTACTGATGATGAACATGCCCGGTTACTGGAACGTTGTGAGGGTAAACAACT
GGCGGTATGGATGCGGCGGGACCAGAGAAAAATCACTCAGGGTCAATGCCAGCGCTTCGTTAATACAGA
TGTAGGTGTTCCACAGGGTAGCCAGCAGCATCCTGCGATGCAGATCCGGAACATAATGGTGCAGGGCGCT
GACTTCCGCGTTTCCAGACTTTACGAAACACGGAAACCGAAGACCATTCATGTTGTTGCTCAGGTCGCAGA
CGTTTTGCAGCAGCAGTCGCTTCACGTTCGCTCGCGTATCGGTGATTCATTCTGCTAACCAGTAAGGCAAC
CCCGCCAGCCTAGCCGGGTCCTCAACGACAGGAGCACGATCATGCGCACCCGTGGGGCCGCCATGCCGG
CGATAATGGCCTGCTTCTCGCCGAAACGTTTGGTGGCGGGACCAGTGACGAAGGCTTGAGCGAGGGCGT

GCAAGATTCCGAATACCGCAAGCGACAGGCCGATCATCGTCGCGCTCCAGCGAAAGCGGTCCTCGCCGAA
AATGACCCAGAGCGCTGCCGGCACCTGTCCTACGAGTTGCATGATAAAGAAGACAGTCATAAGTGCGGCG
ACGATAGTCATGCCCCGCGCCCACCGGAAGGAGCTGACTGGGTTGAAGGCTCTCAAGGGCATCGGTCGA
GATCCCGGTGCCTAATGAGTGAGCTAACTTACATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGG
AAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGCG
CCAGGGTGGTTTTTCTTTTCACCAGTGAGACGGGCAACAGCTGATTGCCCTTCACCGCCTGGCCCTGAGAG
AGTTGCAGCAAGCGGTCCACGCTGGTTTGCCCCAGCAGGCGAAAATCCTGTTTGATGGTGGTTAACGGCG
GGATATAACATGAGCTGTCTTCGGTATCGTCGTATCCCACTACCGAGATATCCGCACCAACGCGCAGCCCG
GACTCGGTAATGGCGCGCATTGCGCCCAGCGCCATCTGATCGTTGGCAACCAGCATCGCAGTGGGAACGA
TGCCCTCATTCAGCATTTGCATGGTTTGTTGAAAACCGGACATGGCACTCCAGTCGCCTTCCCGTTCCGCTA
TCGGCTGAATTTGATTGCGAGTGAGATATTTATGCCAGCCAGCCAGACGCAGACGCGCCGAGACAGAACT
TAATGGGCCCGCTAACAGCGCGATTTGCTGGTGACCCAATGCGACCAGATGCTCCACGCCCAGTCGCGTA
CCGTCTTCATGGGAGAAAATAATACTGTTGATGGGTGTCTGGTCAGAGACATCAAGAAATAACGCCGGAA
CATTAGTGCAGGCAGCTTCCACAGCAATGGCATCCTGGTCATCCAGCGGATAGTTAATGATCAGCCCACTG
ACGCGTTGCGCGAGAAGATTGTGCACCGCCGCTTTACAGGCTTCGACGCCGCTTCGTTCTACCATCGACAC
CACCACGCTGGCACCCAGTTGATCGGCGCGAGATTTAATCGCCGCGACAATTTGCGACGGCGCGTGCAGG
GCCAGACTGGAGGTGGCAACGCCAATCAGCAACGACTGTTTGCCCGCCAGTTGTTGTGCCACGCGGTTGG
GAATGTAATTCAGCTCCGCCATCGCCGCTTCCACTTTTTCCCGCGTTTTCGCAGAAACGTGGCTGGCCTGGT
TCACCACGCGGGAAACGGTCTGATAAGAGACACCGGCATACTCTGCGACATCGTATAACGTTACTGGTTTC
ACATTCACCACCCTGAATTGACTCTCTTCCGGGCGCTATCATGCCATACCGCGAAAGGTTTTGCGCCATTCG
ATGGTGTCCGGGATCTCGACGCTCTCCCTTATGCGACTCCTGCATTAGGAAGCAGCCCAGTAGTAGGTTGA
GGCCGTTGAGCACCGCCGCCGCAAGGAATGGTGCATGCAAGGAGATGGCGCCCAACAGTCCCCCGGCCA
CGGGGCCTGCCACCATACCCACGCCGAAACAAGCGCTCATGAGCCCGAAGTGGCGAGCCCGATCTTCCCC
ATCGGTGATGTCGGCGATATAGGCGCCAGCAACCGCACCTGTGGCGCCGGTGATGCCGGCCACGATGCG
TCCGGCGTAGAGGATCGAGATCTCGATCCCGCGAAATTAATACGACTCACTATAGGGGAATTGTGAGCGG
ATAACAATTCCCCTCTAGAAATAATTTTGTTTAACTTTAAGAAGGAGATATACCATGAAATACCTATTGCCT
ACGGCAGCCGCTGGATTGTTATTACTCGCGGCCCAGCCGGCCATGGCATCTAGGACTCCGGAAATGCCGG
TTCTGGAAAATCGTGCAGCACAGGGTGATATTACCGCACCGGGTGGTGCACGTCGTCTGACCGGTGATCA
GACCGCAGCACTGCGTGATAGCCTGAGCGATAAACCGGCAAAAAACATTATTCTGCTGATTGGTGATGGC
ATGGGTGATAGCGAAATTACCGCAGCACGTAATTATGCCGAAGGTGCCGGTGGTTTTTTTAAAGGTATTG
ATGCACTGCCGCTGACAGGTCAGTATACCCATTATGCACTGAATAAAAAAACCGGCAAACCGGATTATGTT
ACCGATAGCGCAGCAAGCGCAACCGCATGGTCAACCGGTGTTAAAACCTATAATGGTGCACTGGGTGTGG
ATATTCATGAAAAAGATCATCCGACCATTCTGGAAATGGCAAAAGCAGCAGGTCTGGCAACCGGTAATGT
TAGCACCGCAGAACTGCAGGATGCAACACCGGCAGCACTGGTTGCACATGTTACCAGCCGTAAATGTTAT
GGTCCGAGCGCAACCAGCGAAAAATGTCCGGGTAATGCACTGGAAAAAGGTGGTAAAGGTAGCATTACC
GAACAGCTGCTGAATGCACGTGCAGATGTTACCCTGGGTGGTGGTGCAAAAACCTTTGCAGAAACCGCAA
CCGCAGGCGAATGGCAGGGTAAAACCCTGCGTGAACAGGCACAGGCACGTGGTTATCAGCTGGTTAGTG
ATGCAGCAAGCCTGAATAGCGTTACCGAAGCAAATCAGCAGAAACCGCTGCTGGGTCTGTTTGCAGATGG
TAATATGCCGGTTCGTTGGCTGGGTCCGAAAGCAACCTATCATGGTAATATTGATAAACCGGCTGTTACCT
GTACCCCGAATCCGCAGCGTAATGATAGCGTTCCGACCCTGGCACAGATGACCGATAAAGCAATTGAACT
GCTGAGCAAAAATGAAAAAGGCTTTTTTCTGCAGGTTGAAGGTGCCAGCATTGATAAACAGGATCATGCA
GCAAATCCGTGTGGTCAGATTGGTGAAACCGTTGATCTGGATGAAGCAGTTCAGCGTGCACTGGAATTTG
CAAAAAAAGAAGGTAATACCCTGGTTATTGTGACCGCAGATCATGCACATGCAAGCCAGATTGTTGCACC
GGATACCAAAGCACCGGGTCTGACCCAGGCACTGAATACCAAAGATGGTGCAGTTATGGTTATGAGCTAT
GGTAATAGCGAAGAAGATAGCCAGGAACATACCGGTAGCCAGCTGCGTATTGCAGCATATGGTCCGCAT
GCAGCCAATGTTGTTGGTCTGACCGATCAAACCGACCTGTTTTATACCATGAAAGCAGCACTGGGTCTGAA
AGGCGGCGGTGGTAGCAAAGGTGGTGGCTCTCCGAACCCCGGTCCTTGGGAACAGTGTGGTGGTAAATC
GTATGAAGGTCCGACCGCCTGTGCAAGCGGCACCACCTGTCAGGTTCTGCGTGAAACGTTTTCACAGTGT
ATTCCTGGTGGTAAAGGCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGG
AAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTT
GAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGAT

>pelB-AP-Ehux1b2D-pBR1a (6804 bp)

TGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGAC
CGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGC
TTTCCCCGTCAAGCTCTAAATCGGGGGCTCCCTTTAGGGTTCCGATTTAGTGCTTTACGGCACCTCGACCCC
AAAAAACTTGATTAGGGTGATGGTTCACGTAGTGGGCCATCGCCCTGATAGACGGTTTTTCGCCCTTTGAC
GTTGGAGTCCACGTTCTTTAATAGTGGACTCTTGTTCCAAACTGGAACAACACTCAACCCTATCTCGGTCTA
TTCTTTTGATTTATAAGGGATTTTGCCGATTTCGGCCTATTGGTTAAAAAATGAGCTGATTTAACAAAAATT
TAACGCGAATTTTAACAAAATATTAACGCTTACAATTTAGGTGGCACTTTTCGGGGAAATGTGCGCGGAAC
CCCTATTTGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAATTAATTCTTAGAAAAACTCATC
GAGCATCAAATGAAACTGCAATTTATTCATATCAGGATTATCAATACCATATTTTTGAAAAAGCCGTTTCTG
TAATGAAGGAGAAAACTCACCGAGGCAGTTCCATAGGATGGCAAGATCCTGGTATCGGTCTGCGATTCCG
ACTCGTCCAACATCAATACAACCTATTAATTTCCCCTCGTCAAAAATAAGGTTATCAAGTGAGAAATCACCA
TGAGTGACGACTGAATCCGGTGAGAATGGCAAAAGTTTATGCATTTCTTTCCAGACTTGTTCAACAGGCCA
GCCATTACGCTCGTCATCAAAATCACTCGCATCAACCAAACCGTTATTCATTCGTGATTGCGCCTGAGCGA
GACGAAATACGCGATCGCTGTTAAAAGGACAATTACAAACAGGAATCGAATGCAACCGGCGCAGGAACA
CTGCCAGCGCATCAACAATATTTTCACCTGAATCAGGATATTCTTCTAATACCTGGAATGCTGTTTTCCCGG
GGATCGCAGTGGTGAGTAACCATGCATCATCAGGAGTACGGATAAAATGCTTGATGGTCGGAAGAGGCA
TAAATTCCGTCAGCCAGTTTAGTCTGACCATCTCATCTGTAACATCATTGGCAACGCTACCTTTGCCATGTTT
CAGAAACAACTCTGGCGCATCGGGCTTCCCATACAATCGATAGATTGTCGCACCTGATTGCCCGACATTAT
CGCGAGCCCATTTATACCCATATAAATCAGCATCCATGTTGGAATTTAATCGCGGCCTAGAGCAAGACGTT
TCCCGTTGAATATGGCTCATAACACCCCTTGTATTACTGTTTATGTAAGCAGACAGTTTTATTGTTCATGACC
AAAATCCCTTAACGTGAGTTTTCGTTCCACTGAGCGTCAGACCCCGTAGAAAAGATCAAAGGATCTTCTTG
AGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACCAGCGGTGGTTTGTTT
GCCGGATCAAGAGCTACCAACTCTTTTTCCGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACT
GTCCTTCTAGTGTAGCCGTAGTTAGGCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCTCTG
CTAATCCTGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAAGACGATA
GTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGGGTTCGTGCACACAGCCCAGCTTGGAGCGAAC
GACCTACACCGAACTGAGATACCTACAGCGTGAGCTATGAGAAAGCGCCACGCTTCCCGAAGGGAGAAA
GGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGA
AACGCCTGGTATCTTTATAGTCCTGTCGGGTTTCGCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCG
TCAGGGGGGCGGAGCCTATGGAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGC
CTTTTGCTCACATGTTCTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTATTACCGCCTTTGAGTGAGC
TGATACCGCTCGCCGCAGCCGAACGACCGAGCGCAGCGAGTCAGTGAGCGAGGAAGCGGAAGAGCGCC
TGATGCGGTATTTTCTCCTTACGCATCTGTGCGGTATTTCACACCGCATATATGGTGCACTCTCAGTACAAT
CTGCTCTGATGCCGCATAGTTAAGCCAGTATACACTCCGCTATCGCTACGTGACTGGGTCATGGCTGCGCC
CCGACACCCGCCAACACCCGCTGACGCGCCCTGACGGGCTTGTCTGCTCCCGGCATCCGCTTACAGACAAG
CTGTGACCGTCTCCGGGAGCTGCATGTGTCAGAGGTTTTCACCGTCATCACCGAAACGCGCGAGGCAGCT
GCGGTAAAGCTCATCAGCGTGGTCGTGAAGCGATTCACAGATGTCTGCCTGTTCATCCGCGTCCAGCTCGT
TGAGTTTCTCCAGAAGCGTTAATGTCTGGCTTCTGATAAAGCGGGCCATGTTAAGGGCGGTTTTTTCCTGT
TTGGTCACTGATGCCTCCGTGTAAGGGGGATTTCTGTTCATGGGGGTAATGATACCGATGAAACGAGAGA
GGATGCTCACGATACGGGTTACTGATGATGAACATGCCCGGTTACTGGAACGTTGTGAGGGTAAACAACT
GGCGGTATGGATGCGGCGGGACCAGAGAAAAATCACTCAGGGTCAATGCCAGCGCTTCGTTAATACAGA
TGTAGGTGTTCCACAGGGTAGCCAGCAGCATCCTGCGATGCAGATCCGGAACATAATGGTGCAGGGCGCT
GACTTCCGCGTTTCCAGACTTTACGAAACACGGAAACCGAAGACCATTCATGTTGTTGCTCAGGTCGCAGA
CGTTTTGCAGCAGCAGTCGCTTCACGTTCGCTCGCGTATCGGTGATTCATTCTGCTAACCAGTAAGGCAAC
CCCGCCAGCCTAGCCGGGTCCTCAACGACAGGAGCACGATCATGCGCACCCGTGGGGCCGCCATGCCGG
CGATAATGGCCTGCTTCTCGCCGAAACGTTTGGTGGCGGGACCAGTGACGAAGGCTTGAGCGAGGGCGT
GCAAGATTCCGAATACCGCAAGCGACAGGCCGATCATCGTCGCGCTCCAGCGAAAGCGGTCCTCGCCGAA
AATGACCCAGAGCGCTGCCGGCACCTGTCCTACGAGTTGCATGATAAAGAAGACAGTCATAAGTGCGGCG
ACGATAGTCATGCCCCGCGCCCACCGGAAGGAGCTGACTGGGTTGAAGGCTCTCAAGGGCATCGGTCGA
GATCCCGGTGCCTAATGAGTGAGCTAACTTACATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGG
AAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGCG
CCAGGGTGGTTTTTCTTTTCACCAGTGAGACGGGCAACAGCTGATTGCCCTTCACCGCCTGGCCCTGAGAG
AGTTGCAGCAAGCGGTCCACGCTGGTTTGCCCCAGCAGGCGAAAATCCTGTTTGATGGTGGTTAACGGCG
GGATATAACATGAGCTGTCTTCGGTATCGTCGTATCCCACTACCGAGATATCCGCACCAACGCGCAGCCCG

GACTCGGTAATGGCGCGCATTGCGCCCAGCGCCATCTGATCGTTGGCAACCAGCATCGCAGTGGGAACGA
TGCCCTCATTCAGCATTTGCATGGTTTGTTGAAAACCGGACATGGCACTCCAGTCGCCTTCCCGTTCCGCTA
TCGGCTGAATTTGATTGCGAGTGAGATATTTATGCCAGCCAGCCAGACGCAGACGCGCCGAGACAGAACT
TAATGGGCCCGCTAACAGCGCGATTTGCTGGTGACCCAATGCGACCAGATGCTCCACGCCCAGTCGCGTA
CCGTCTTCATGGGAGAAAATAATACTGTTGATGGGTGTCTGGTCAGAGACATCAAGAAATAACGCCGGAA
CATTAGTGCAGGCAGCTTCCACAGCAATGGCATCCTGGTCATCCAGCGGATAGTTAATGATCAGCCCACTG
ACGCGTTGCGCGAGAAGATTGTGCACCGCCGCTTTACAGGCTTCGACGCCGCTTCGTTCTACCATCGACAC
CACCACGCTGGCACCCAGTTGATCGGCGCGAGATTTAATCGCCGCGACAATTTGCGACGGCGCGTGCAGG
GCCAGACTGGAGGTGGCAACGCCAATCAGCAACGACTGTTTGCCCGCCAGTTGTTGTGCCACGCGGTTGG
GAATGTAATTCAGCTCCGCCATCGCCGCTTCCACTTTTTCCCGCGTTTTCGCAGAAACGTGGCTGGCCTGGT
TCACCACGCGGGAAACGGTCTGATAAGAGACACCGGCATACTCTGCGACATCGTATAACGTTACTGGTTTC
ACATTCACCACCCTGAATTGACTCTCTTCCGGGCGCTATCATGCCATACCGCGAAAGGTTTTGCGCCATTCG
ATGGTGTCCGGGATCTCGACGCTCTCCCTTATGCGACTCCTGCATTAGGAAGCAGCCCAGTAGTAGGTTGA
GGCCGTTGAGCACCGCCGCCGCAAGGAATGGTGCATGCAAGGAGATGGCGCCCAACAGTCCCCCGGCCA
CGGGGCCTGCCACCATACCCACGCCGAAACAAGCGCTCATGAGCCCGAAGTGGCGAGCCCGATCTTCCCC
ATCGGTGATGTCGGCGATATAGGCGCCAGCAACCGCACCTGTGGCGCCGGTGATGCCGGCCACGATGCG
TCCGGCGTAGAGGATCGAGATCTCGATCCCGCGAAATTAATACGACTCACTATAGGGGAATTGTGAGCGG
ATAACAATTCCCCTCTAGAAATAATTTTGTTTAACTTTAAGAAGGAGATATACCATGAAATACCTATTGCCT
ACGGCAGCCGCTGGATTGTTATTACTCGCGGCCCAGCCGGCCATGGCATCTAGGACTCCGGAAATGCCGG
TTCTGGAAAATCGTGCAGCACAGGGTGATATTACCGCACCGGGTGGTGCACGTCGTCTGACCGGTGATCA
GACCGCAGCACTGCGTGATAGCCTGAGCGATAAACCGGCAAAAAACATTATTCTGCTGATTGGTGATGGC
ATGGGTGATAGCGAAATTACCGCAGCACGTAATTATGCCGAAGGTGCCGGTGGTTTTTTTAAAGGTATTG
ATGCACTGCCGCTGACAGGTCAGTATACCCATTATGCACTGAATAAAAAAACCGGCAAACCGGATTATGTT
ACCGATAGCGCAGCAAGCGCAACCGCATGGTCAACCGGTGTTAAAACCTATAATGGTGCACTGGGTGTGG
ATATTCATGAAAAAGATCATCCGACCATTCTGGAAATGGCAAAAGCAGCAGGTCTGGCAACCGGTAATGT
TAGCACCGCAGAACTGCAGGATGCAACACCGGCAGCACTGGTTGCACATGTTACCAGCCGTAAATGTTAT
GGTCCGAGCGCAACCAGCGAAAAATGTCCGGGTAATGCACTGGAAAAAGGTGGTAAAGGTAGCATTACC
GAACAGCTGCTGAATGCACGTGCAGATGTTACCCTGGGTGGTGGTGCAAAAACCTTTGCAGAAACCGCAA
CCGCAGGCGAATGGCAGGGTAAAACCCTGCGTGAACAGGCACAGGCACGTGGTTATCAGCTGGTTAGTG
ATGCAGCAAGCCTGAATAGCGTTACCGAAGCAAATCAGCAGAAACCGCTGCTGGGTCTGTTTGCAGATGG
TAATATGCCGGTTCGTTGGCTGGGTCCGAAAGCAACCTATCATGGTAATATTGATAAACCGGCTGTTACCT
GTACCCCGAATCCGCAGCGTAATGATAGCGTTCCGACCCTGGCACAGATGACCGATAAAGCAATTGAACT
GCTGAGCAAAAATGAAAAAGGCTTTTTTCTGCAGGTTGAAGGTGCCAGCATTGATAAACAGGATCATGCA
GCAAATCCGTGTGGTCAGATTGGTGAAACCGTTGATCTGGATGAAGCAGTTCAGCGTGCACTGGAATTTG
CAAAAAAAGAAGGTAATACCCTGGTTATTGTGACCGCAGATCATGCACATGCAAGCCAGATTGTTGCACC
GGATACCAAAGCACCGGGTCTGACCCAGGCACTGAATACCAAAGATGGTGCAGTTATGGTTATGAGCTAT
GGTAATAGCGAAGAAGATAGCCAGGAACATACCGGTAGCCAGCTGCGTATTGCAGCATATGGTCCGCAT
GCAGCCAATGTTGTTGGTCTGACCGATCAAACCGACCTGTTTTATACCATGAAAGCAGCACTGGGTCTGAA
AGGCGGCGGTGGTAGCAAAGGTGGTGGCTCTCCGAACCCCGGTCCTTGGGAACAGTGTGGTGGTAAATC
GTATGAAGGTCCGACCGCCTGTCCCCGTGAATACACCTGTCAGTATCGCAATCCGTATTATTCACAGTGTCT
GCCTGGTGGTAAAGGCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGA
AGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTG
AGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGAT

CyDisCo plasmid

>CyDisCo plasmid pMJS205 (7148 bp)

GAATTCCGGATGAGCATTCATCAGGCGGGCAAGAATGTGAATAAAGGCCGGATAAAACTTGTGCTTATTT
TTCTTTACGGTCTTTAAAAAGGCCGTAATATCCAGCTGAACGGTCTGGTTATAGGTACATTGAGCAACTGA
CTGAAATGCCTCAAAATGTTCTTTACGATGCCATTGGGATATATCAACGGTGGTATATCCAGTGATTTTTTT
CTCCATTTTAGCTTCCTTAGCTCCTGAAAATCTCGATAACTCAAAAAATACGCCCGGTAGTGATCTTATTTCA
TTATGGTGAAAGTTGGAACCTCTTACGTGCCGATCAACGTCTCATTTTCGCCAAAAGTTGGCCCAGGGCTT
CCCGGTATCAACAGGGACACCAGGATTTATTTATTCTGCGAAGTGATCTTCCGTCACAGGTATTTATTCGG

CGCAAAGTGCGTCGGGTGATGCTGCCAACTTACTGATTTAGTGTATGATGGTGTTTTTGAGGTGCTCCAGT
GGCTTCTGTTTCTATCAGCTGTCCCTCCTGTTCAGCTACTGACGGGGTGGTGCGTAACGGCAAAAGCACCG
CCGGACATCAGCGCTAGCGGAGTGTATACTGGCTTACTATGTTGGCACTGATGAGGGTGTCAGTGAAGTG
CTTCATGTGGCAGGAGAAAAAAGGCTGCACCGGTGCGTCAGCAGAATATGTGATACAGGATATATTCCGC
TTCCTCGCTCACTGACTCGCTACGCTCGGTCGTTCGACTGCGGCGAGCGGAAATGGCTTACGAACGGGGC
GGAGATTTCCTGGAAGATGCCAGGAAGATACTTAACAGGGAAGTGAGAGGGCCGCGGCAAAGCCGTTTT
TCCATAGGCTCCGCCCCCCTGACAAGCATCACGAAATCTGACGCTCAAATCAGTGGTGGCGAAACCCGAC
AGGACTATAAAGATACCAGGCGTTTCCCCTGGCGGCTCCCTCGTGCGCTCTCCTGTTCCTGCCTTTCGGTTT
ACCGGTGTCATTCCGCTGTTATGGCCGCGTTTGTCTCATTCCACGCCTGACACTCAGTTCCGGGTAGGCAG
TTCGCTCCAAGCTGGACTGTATGCACGAACCCCCCGTTCAGTCCGACCGCTGCGCCTTATCCGGTAACTATC
GTCTTGAGTCCAACCCGGAAAGACATGCAAAAGCACCACTGGCAGCAGCCACTGGTAATTGATTTAGAGG
AGTTAGTCTTGAAGTCATGCGCCGGTTAAGGCTAAACTGAAAGGACAAGTTTTGGTGACTGCGCTCCTCCA
AGCCAGTTACCTCGGTTCAAAGAGTTGGTAGCTCAGAGAACCTTCGAAAAACCGCCCTGCAAGGCGGTTT
TTTCGTTTTCAGAGCAAGAGATTACGCGCAGACCAAAACGATCTCAAGAAGATCATCTTATTAATCAGATA
AAATATTTCTAGATTTCAGTGCAATTTATCTCTTCAAATGTAGCACCTGAAGTCAGCCCCATACGATATAAG
TTGTAATTCTCATGTTTGACAGCTTATCATCGATAAGCTTTAATGCGGTAGTTTATCACAGTTAAATTGCTA
ACGCAGTCAGGCACCGTGTATGAAATCTAACAATGCGCTCATCGTCATCCTCGGCACCGTCACCCTGGATG
CTGTAGGCATAGGCTTGGTTATGCCGGTACTGCCGGGCCTCTTGCGGGATATCGTCCATTCCGACAGCATC
GCCAGTCACTATGGCGTGCTGCTAGCGCTATATGCGTTGATGCAATTTCTATGCGCACCCGTTCTCGGAGC
ACTGTCCGACCGCTTTGGCCGCCGCCCAGTCCTGCTCGCTTCGCTACTTGGAGCCACTATCGACTACGCGA
TCATGGCGACCACACCCGTCCTGTGGATCCGGCCCATTGGCTGCCTCCCACACTTGGATATGCCTCCTCGG
AGCCTTATAGAATTGTTTATAAGACTTGCGCATTATTTGACCTCCAATGCGAACAAAGGGAAACCGCTGTG
GTCTCCCTTTAGTGAGTTCAATTAATTATCCACGGTCAGAAGTGACCAGTTCGTTCTTCTCCCACCAACGCT
TAAGGTCGAACGAAGGGCAAGCCTTCGGCGCCACCTCATGATGGGCGCGAAGACCAGCGCCTTCGTACTT
AGCCAGCAGTGTGACAAGCAGTGAGCGAAGGGATTGCATTTGGGCTGGCGTAAAGTTAGCGTCGAACTT
ACCTTTATCGTCGATACCACCAACAAGGCAGACGCCGATAGAGTTGTGGTTGTAACCCTTAGCGTGAGAG
CCTACAGCCATCTCATCTCGTCCTGCCTCCACAGTACCGTCTCGCTTGATGATAAAGTGGTATCCCACATCG
AGCCAACCCTGCTCTTTGTGCCACTGGCGAATCTCACGGACACCAACATTCTGACTTGGCTTGGTAGCCGA
GCAGTGAACAAAGATTGCGTCAGTAGATTCACGTTGTTTAAACTGTACACGAGCCATTATTTCTTTCCTCCT
TTCCTTTTTAATCTATCAAAGGGGACCCGGATCCTCTACGCCGGACGCATCGTGGCCGGCATCACCGGCGC
CACAGGTGCGGTTGCTGGCGCCTATATCGCCGACATCACCGATGGGGAAGATCGGGCTCGCCACTTCGGG
CTCATGAGCGCTTGTTTCGGCGTGGGTATGGTGGCAGGCCCCGTGGCCGGGGGACTGTTGGGCGCCATCT
CCTTGCATGCACCATTCCTTGCGGCGGCGGTGCTCAACGGCCTCAACCTACTACTGGGCTGCTTCCTAATG
CAGGAGTCGCATAAGGGAGAGCGTCGACCGATGCCCTTGAGAGCCTTCAACCCAGTCAGCTCCTTCCGGT
GGGCGCGGGGCATGACTATCGTCGCCGCACTTATGACTGTCTTCTTTATCATGCAACTCGTAGGACAGGTG
CCGGCAGCGCTCTGGGTCATTTTCGGCGAGGACCGCTTTCGCTGGAGCGCGACGATGATCGGCCTGTCGC
TTGCGGTATTCGGAATCTTGCACGCCCTCGCTCAAGCCTTCGTCACTGGTCCCGCCACCAAACGTTTCGGC
GAGAAGCAGGCCATTATCGCATGCATCCTGGCGCCCAATACGCAAACCGCCTCTCCCCGCGCGTTGGCCG
ATTCATTAATGCAGCTGGCACGACAGGTTTCCCGACTGGAAAGCGGGCAGTGAGCGCAACGCAATTAATG
TAAGTTAGCTCACTCATTAGGCACAATTCTCATGTTTGACAGCTTATCATCGACTGCACGGTGCACCAATGC
TTCTGGCGTCAGGCAGCCATCGGAAGCTGTGGTATGGCTGTGCAGGTCGTAAATCACTGCATAATTCGTG
TCGCTCAAGGCGCACTCCCGTTCTGGATAATGTTTTTTGCGCCGACATCATAACGGTTCTGGCAAATATTCT
GAAATGAGCTGTTGACAATTAATCATCGGCTCGTATAATGTGTGGAATTGTGAGCGGATAACAATTTCACA
CAGGAAACAGCCAGTCCGTTTAGGTGTTTTCACGAGCTCTAGAAATAATTTTGTTTAACTTTAAGAAGGAG
ATACATATGAAAGCCATCGATAAAATGACCGACAATCCGCCTCAAGAGGGTCTGAGTGGTCGTAAAATCA
TCTATGATGAGGACGGCAAACCATGTCGCAGCTGCAATACCCTGCTGGACTTCCAGTATGTTACCGGGAA
AATCTCAAATGGCCTGAAAAACCTGAGCAGCAATGGTAAACTGGCGGGTACAGGTGCTCTGACTGGGGA
AGCATCTGAACTGATGCCGGGTTCTCGTACCTATCGTAAAGTCGATCCTCCGGATGTTGAACAGCTGGGAC
GTTCTTCATGGACACTGCTGCATAGCGTAGCAGCCTCTTATCCAGCTCAACCAACCGATCAGCAAAAAGGC
GAAATGAAACAGTTCCTGAACATCTTCAGCCACATCTATCCGTGTAACTGGTGCGCTAAAGACTTCGAAAA
ATATATCCGTGAGAATGCCCCTCAAGTTGAATCACGTGAGGAGCTGGGTCGCTGGATGTGTGAGGCCCAT
AACAAAGTGAACAAAAAGCTGCGCAAACCGAAGTTCGACTGTAACTTCTGGGAAAAACGCTGGAAAGAT
GGTTGGGATGAGTAATAAGGATCCGAATTCTCTAGTAATAATTTTGTTTAACTTTAAGAAGGAGATACATA
TGGATGCCCCAGAGGAGGAAGATCACGTCCTGGTTCTGCGTAAAAGCAACTTCGCTGAAGCACTGGCAGC
TCACAAATATCTGCTGGTCGAGTTCTATGCTCCGTGGTGTGGTCATTGCAAAGCCCTGGCTCCGGAATATG

CTAAAGCAGCCGGCAAACTGAAAGCTGAGGGCAGTGAAATTCGTCTGGCCAAAGTGGACGCTACCGAAG
AATCAGATCTGGCACAACAGTATGGTGTTCGTGGTTATCCGACTATCAAATTTTTCCGTAACGGCGATACA
GCAAGCCCTAAAGAGTATACCGCTGGCCGTGAAGCTGATGATATCGTGAACTGGCTGAAAAAACGTACAG
GTCCGGCGGCAACGACTCTGCCTGATGGTGCCGCTGCCGAGTCACTGGTAGAATCATCCGAAGTGGCCGT
GATTGGCTTCTTTAAAGACGTGGAGAGCGATTCAGCAAAACAGTTCCTGCAAGCAGCTGAAGCGATTGAT
GACATCCCGTTTGGTATTACGAGCAATAGCGACGTGTTCTCCAAATATCAACTGGACAAAGACGGTGTGG
TTCTGTTCAAAAAATTCGACGAAGGCCGTAACAACTTTGAAGGTGAGGTGACCAAAGAAAACCTGCTGGA
CTTTATCAAACACAATCAACTGCCGCTGGTGATTGAGTTCACCGAACAGACAGCTCCGAAAATCTTTGGCG
GCGAGATCAAAACCCACATTCTGCTGTTTCTGCCTAAAAGTGTGTCCGACTATGACGGCAAACTGAGCAAC
TTCAAAACCGCCGCTGAGAGCTTTAAAGGCAAAATCCTGTTCATCTTCATCGACAGCGATCATACCGACAA
CCAGCGTATTCTGGAGTTTTTTGGCCTGAAAAAAGAGGAATGTCCGGCCGTTCGCCTGATTACACTGGAA
GAGGAGATGACGAAATATAAACCAGAAAGCGAGGAGCTGACAGCTGAACGTATTACCGAGTTCTGCCAC
CGTTTTCTGGAAGGCAAAATCAAACCTCATCTGATGTCCCAAGAACTGCCAGAAGATTGGGATAAACAGC
CTGTGAAAGTGCTGGTAGGCAAAAACTTCGAGGATGTGGCGTTCGACGAGAAAAAAAACGTGTTTGTGG
AGTTTTATGCCCCTTGGTGTGGACACTGCAAACAGCTGGCACCGATTTGGGATAAACTGGGCGAAACCTA
TAAAGATCATGAAAACATTGTTATTGCCAAAATGGACAGCACCGCCAATGAAGTCGAAGCCGTGAAAGTT
CATTCGTTTCCGACCCTGAAATTCTTTCCTGCCAGCGCCGATCGTACTGTGATCGATTATAACGGGGAGCG
TACCCTGGATGGTTTTAAAAAATTCCTGGAGAGCGGTGGTCAAGATGGTGCCGGTGATGATGACGATCTG
GAGGATCTGGAAGAGGCTGAAGAACCGGATATGGAGGAGGATGACGATCAGAAAGCAGTCAAAGACGA
GCTGTGATAAGGATCCGAATTCACTAGTGAGCTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACCACCA
CCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAG
CAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCCTAGGCACGGGTGCG
CATGATCGTGCTCCTGTCGTTGAGGACCCGGCTAGGCTGGCGGGGTTGCCTTACTGGTTAGCAGAATGAA
TCACCGATACGCGAGCGAACGTGAAGCGACTGCTGCTGCAAAACGTCTGCGACCTGAGCAACAACATGAA
TGGTCTTCGGTTTCCGTGTTTCGTAAAGTCTGGAAACGCGGAAGTCCCCTACGTGCTGCTGAAGTTGCCCG
CAACAGAGAGTGGAACCAACCGGTGATACCACGATACTATGACTGAGAGTCAACGCCATGAGCGGCCTCA
TTTCTTATTCTGAGTTACAACAGTCCGCACCGCTGTCCGGTAGCTCCTTCCGGTGGGCGCGGGGCATGACT
ATCGTCGCCGCACTTATGACTGTCTTCTTTATCATGCAACTCGTAGGACAGGTGCCGGCAGCGCCCAACAG
TCCCCCGGCCACGGGGCCTGCCACCATACCCACGCCGAAACAAGCGCCCTGCACCATTATGTTCCGGATCT
GCATCGCAGGATGCTGCTGGCTACCCTGTGGAACACCTACATCTGTATTAACGAAGCGCTAACCGTTTTTA
TCAGGCTCTGGGAGGCAGAATAAATGATCATATCGTCAATTATTACCTCCACGGGGAGAGCCTGAGCAAA
CTGGCCTCAGGCATTTGAGAAGCACACGGTCACACTGCTTCCGGTAGTCAATAAACCGGTAAACCAGCAA
TAGACATAAGCGGCTATTTAACGACCCTGCCCTGAACCGACGACCGGGTCGAATTTGCTTTCGAATTTCTG
CCATTCATCCGCTTATTATCACTTATTCAGGCGTAGCACCAGGCGTTTAAGGGCACCAATAACTGCCTTAAA
AAAATTACGCCCCGCCCTGCCACTCATCGCAGTACTGTTGTAATTCATTAAGCATTCTGCCGACATGGAAG
CCATCACAGACGGCATGATGAACCTGAATCGCCAGCGGCATCAGCACCTTGTCGCCTTGCGTATAATATTT
GCCCATGGTGAAAACGGGGGCGAAGAAGTTGTCCATATTGGCCACGTTTAAATCAAAACTGGTGAAACTC
ACCCAGGGATTGGCTGAGACGAAAAACATATTCTCAATAAACCCTTTAGGGAAATAGGCCAGGTTTTCACC
GTAACACGCCACATCTTGCGAATATATGTGTAGAAACTGCCGGAAATCGTCGTGGTATTCACTCCAGAGCG
ATGAAAACGTTTCAGTTTGCTCATGGAAAACGGTGTAACAAGGGTGAACACTATCCCATATCACCAGCTCA
CCGTCTTTCATTGCCATACG

CBM1 protein sequences in multiple sequence alignment

GenBank accession numbers of the sequences aligned in Figure 21.

1. **Cel7A** [*Trichoderma reesei*] (this study)
2. endoglucanase 1 [*Penicillium oxalicum*] - AGW24293.1
3. xylanase/cellobiohydrolase [*Talaromyces funiculosus*] - CAC85737.1
4. cellobiohydrolase I, partial [*Geotrichum candidum*] - AIO10971.1
5. endoglucanase I, partial [*Trichoderma longibrachiatum*] - AEC03714.1
6. cellobiohydrolase I Cel7A [*Talaromyces cellulolyticus*] - GAM33347.1
7. cellulase [*Irpex lacteus*] - BAA76365.1|
8. cellobiohydrolase family protein 61, partial [*Chaetomium thermophilum*] - AGY80103.1
9. cellobiohydrolase I [*Penicillium granulatum*] - AGU16949.1
10. Cel7A, partial [*Aspergillus fischeri*] - ALE19913.1
11. cellobiohydrolase I [*Alternaria japonica*] - AGU16948.1
12. cellobiohydrolase B [*Aspergillus niger*] - AKH61141.1
13. cellobiohydrolase [*Aspergillus terreus*] - AAW68437.2
14. cellobiohydrolase [*Penicillium oxalicum*] - AKI32221.1
15. cellobiohydrolase I [*Penicillium oxalicum*] - ALO81607.1
16. cellobiohydrolase 2 [*Penicillium oxalicum*] - AGW24292.1
17. exo-cellobiohydrolase [*Penicillium oxalicum*] - AEF33951.1
18. cellobiohydrolase I [*Chaetomium murorum*] - AGU16947.1
19. 1,4-beta-D-glucan cellobiohydrolase B precursor [*Aspergillus niger*] - AEF58998.1
20. cellobiohydrolase family protein 45, partial [*Chaetomium thermophilum*] - AGY80101.1
21. xylanase 4 [*Penicillium oxalicum*] - AGW24301.1
22. Xylanase B [*Neocallimastix patriciarum*] - AAB30669.1
23. Cellulase [*Humicola grisea var. thermoidea*] - BAA09785.1
24. cellobiohydrolase I, partial [*Penicillium canescens*] - AIL95870.1

25. chitinase [*Trichoderma virens*] - ACI96032.1
26. chitinase [*Beauveria bassiana*] - AIT18883.1
27. putative chitinase [*Metarhizium anisopliae*] - AAY34347.1
28. **Ehux1b2** (R1C3S1-CBM1-2) [*Emiliania huxleyi*] (this study)
29. chitinase chi18-17 [*Trichoderma citrinoviride*] - ADF57311.1